YHTEISKUNTATIETEELLISEN TIETOARKISTON ASIAKKAIDEN TIEDONHAUT: LOKI-TUTKIMUS AILA-AINEISTOPORTAALISTA

Anna Tuominen

Tampereen yliopisto
Informaatiotieteiden yksikkö
Informaatiotutkimus ja interaktiivinen media
Pro gradu -tutkielma
Huhtikuu 2015

TAMPEREEN YLIOPISTO
Informaatiotieteiden yksikkö
Informaatiotutkimus ja interaktiivinen media
TUOMINEN ANNA
Yhteiskuntatieteellisen tietoarkiston asiakkaiden tiedonhaut:
lokianalyysi Aila-aineistoportaalista
Pro gradu -tutkielma
Huhtikuu 2015

Työssä on tarkasteltu Yhteiskuntatieteellisen tietoarkiston asiakkaiden *Aila* -aineistoportaalissa tekemiä hakuja palvelinlokista saadun aineiston valossa. Tuloksista selviävät eri hakumahdollisuuksien käytön osuudet hakujen kokonaismäärästä. Selvästi yleisintä oli yleishaku ja yleishaun hakukentistä vapaasanahaku.

Yleishaun aiheenmukaisista hauista tutkittiin erityisesti asiasanahakukentässä käytettyjen hakutermien vastaavuutta kuvailussa käytettyihin asiasanastoihin sekä vapaasanahauista poimitun otoksen hakutermien tuloksellisuutta. Asiasanahauissa käytetyistä termeistä 60 % oli tietoarkiston käyttämien asiasanastojen asiasanoja.

Vapaasanahaun hakutermeistä poimitun otoksen termeillä toistettiin haut *Ailassa* ja tarkasteltiin niiden tuottamia saanteja. Asiasanastoista löytyvillä termeillä tehtyjä vapaasanahakuja toistettiin sekä vapaasana- että asiasanahakuina ja verrattiin saantien suuruuksia. Lisäksi arvioitiin vapaasanahakujen otoksessa olleita nollatuloksen tuottaneita termejä sekä asiasanahaussa käytettyjä asiasanastoa vastaamattomia termejä. Asiasanahakujen tarkastelu antaa joitain viitteitä myös kuvailussa käytettyjen dokumentaatiokielten, erityisesti Yleisen suomalaisen asiasanaston YSAn, laadusta ja soveltuvuudesta yhteiskuntatieteellisten tutkimusaineistojen kontekstiin.

Tulokset ovat varsin yhdenmukaisia työn taustakirjallisuutena käytetyn, lähinnä tieteellisissä kirjastoissa toteutettujen hakutapoja selvittäneiden lokitutkimusten kanssa. Niissä on havaittu asiakkaiden suosivan yksinkertaisia hakutapoja ja toimivien hakutermien valinnan tuottavan usein vaikeuksia aiheenmukaisessa haussa. Myös havainto vapaasanahaun yleisyydestä on hyvin yhdenmukainen aiempien tutkimusten kanssa.

Avainsanat: lokianalyysi, aineistoarkistot, Yhteiskuntatieteellinen tietoarkisto, metatieto, sisällönkuvailu

1 JOHDANTO

Kansainvälisen kirjastojärjestön IFLAn informaatioammattilaisille laatimien ammattieettisten ohjeiden mukaan tiedon löydettävyys on myös ammattieettinen kysymys:

Librarians and other information workers organize and present content in a way that allows an autonomous user to find the information s/he needs.
*(*IFLA 2015)

Sain kirjasto- ja informaatioalan pätevyyden vuonna 2011 Tampereen yliopiston järjestämästä täydennyskoulutuksesta. Ensimmäisessä harjoittelussani asiasanoitin Yleistä suomalaista asiasanastoa YSAa käyttäen naistutkimuksen ja sukupuolentutkimuksen alan väitöskirjoja. Kohtasin ja pohdiskelin siis tiedon organisoinnin ja löydettävyyden kysymyksiä jo informaatioalan ammattiin opiskellessani. Sen lisäksi, että valitsemani asiasanoituksen olisi osuvasti ja mahdollisimman tyhjentävästi kuvailtava sisältöä, jouduin usein miettimään miten toimisin kun kulloisenkin dokumentin sisältöä kuvaavia termejä ei ollutkaan asiasanastossa. Tämän lisäksi koetin arvailla, minkälaisia termejä asiakkaat mahtavat hakiessaan käyttää.

Alan ammattipätevyyden saatuani työskentelin erilaisissa kirjastoissa, joissa sekä tuotin metatietoa että opetin ja tein tiedonhakua. Sain hiljalleen lisää käsitystä siitä, miten hakua palveleva metatieto toimii tai ei toimi. Asiakkaiden kysymykset antoivat viitettä siitä, miten he osaavat, jos osaavat, hyödyntää asiasanoitusta ja muuta dokumenttien sisältöä kuvailevaa metatietoa tiedonhaussa.

Informaatiotutkimuksen opintoihin paluun myötä tämä opinnäytetutkimus on antanut minulle mahdollisuuden tutustua suoremmin ja tarkemmin siihen, millaisia hakuja asiakkaat eräässä tietojärjestelmässä tekevät. Yhteiskuntatieteellinen tietoarkisto antoi tutkittavakseni tiedonhakuportaalinsa *Ailan* lokiaineiston, josta pääsin suoraan katsomaan asiakkaiden valitsemia hakutapoja ja -termejä. Tämän tutkimukseni tutkimuskysymykset koskivat ensinnäkin eri hakumahdollisuuksien käyttöä ja toisekseen asiasana- ja vapaasanahaussa käytettyjä termejä. Aiempiin arvailuihini ja asiakkaiden puheista saamiini viitteisiin verrattuna sain tämän työn myötä paljon täsmällistä tietoa asiakkaiden tekemästä tiedonhausta ja dokumenttien löydettävyydestä.

Pyörän keksimisestä ei ollut kyse. Tutkimukseni perustaksi lukemani kansainvälinen tutkimus on osoittanut monien havaitsemieni ilmiöiden näkyneen muissakin tiedonhakujärjestelmissä lokeja

tutkittaessa. Olen hyödyntänyt erityisesti kirjastojen avointen aineistoluetteloiden (*Open Access Catalogue, OPAC*) lokeihin perustuvia tutkimuksia, joiden päämääränä on usein ollut kartoittaa asiakkaiden kohtaamia tiedonhakuongelmia ja havaintojen pohjalta ehdottaa parannuksia järjestelmiin ja/tai asiakkaiden opastukseen.

Sekä katsaukseni aiempaan tutkimukseen että oman työni tulokset ovat kuitenkin vakuuttaneet minut siitä, että lokiaineistoista saatua tietoa asiakkaiden tiedonhausta kannattaisi hyödyntää nykyistä enemmän – sekä tiedonhakukäyttäytymisen tarkasteluun yleisesti että erityisesti aiheenmukaiseen hakuun ja sisältöä kuvaavan metatiedon toimivuuteen paneutuen. Lokiaineistot tarjoavat runsaasti aineksia sekä tiedonhaun teorioiden että käytännön tiedonhakujärjestelmien kehittämiseen. Oma tutkimukseni liittyy jälkimmäiseen: esitän sen lopuksi myös muutamia ehdotuksia *Aila* -portaalin hakuominaisuuksien parantamiseksi. Tutkimukseni voi näin tarjota Yhteiskuntatieteelliselle tietoarkistolle käyttökelpoista tietoa, jonka avulla käytännön sovelluksissa voidaan parantaa tiedonhakujärjestelmän toimivuutta. Informaatiotutkimuksen alan opinnäytteenä tutkimukseni aihe on kiinnostava tiedonhakukäyttäytymisen tutkimuksen kannalta yleisesti ja erityisesti sillä voi osaltaan olla tarjottavaa pohdintoihin aiheita kuvailevan metatiedon merkityksestä tiedonhaussa. Arkistoihin liittyvää tiedonhaun tutkimusta on Suomessa tehty vasta vähän, mutta huomionarvoisia tutkimusaiheita on arkistomaailmassa paljonkin.

Entä miksi tutkia juuri Yhteiskuntatieteellisen tietoarkiston aineistojen eli tutkimusaineistojen, hakua? Tutkimusaineistojen julkisesta tallentamisesta on Suomessa viime vuosina keskusteltu paljon ja siitä on tullut olennainen osa julkilausuttuja tutkimuspoliittisia päämääriä niin OECD:ssä (ks. esim. Borg & Kuula 2007, 7–8), Euroopan Unionissa (Euroopan Unioni 2012) kuin Suomen valtiossa (Valtiovarainministeriö 2014). Avoin data -politiikka on nykyisin mukana esimerkiksi Suomen Akatemian tutkimusrahoitusta ohjaavissa periaatteissa ja kaikkien alojen tutkimusrahoitushakemuksilta edellytetään suunnitelmaa aineistonhallinnasta (Borg 2011, 1). Niin kutsutun *tutkimuksen tietoinfrastruktuurin* kehittämiseen on liittynyt muun muassa Opetus- ja kulttuuriministeriön *Tutkimuksen tietoaineistot* -hanke vuosina 2011–2013, johon myös Yhteiskuntatieteellinen tietoarkisto osallistui. Sen tavoitteisiin kuului entistä yhdenmukaisempien tallennus - ja metatietokäytäntöjen edistäminen (Syväjärvi 2014, 4). Nyt käynnissä on ministeriön hanke *Avoin tiede ja tutkimus 2014–2017*. Yhteiskuntatieteellinen tietoarkisto on mukana hankkeessa ja sen *Aila*-portaali on listattuna hankkeen tarjoamien palveluiden joukkoon (Opetus- ja kulttuuriministeriö 2015).

Avoin data -politiikan ja lisääntyvän tutkimusaineistojen tallentamisen ja uudelleenkäytön myötä lisääntyy aineistoarkistoissa varmasti myös monipuolisen informaatioalan osaamisen tarve. Lisääntyvä itsepalvelu asettaa myös uusia vaatimuksia sekä tietojärjestelmille että metatiedolle. Sekä tiedon organisoinnin, tietokantojen tutkimuksen ja suunnittelun että tiedonhakukäyttäytymisen tutkimuksen alueilla onkin aihetta suunnata entistä enemmän mielenkiintoa aineistoarkistoihin.

Tutkimusaineistojen sisällönkuvailussa ja löydettävyydessä on ominaispiirteensä, jotka erottavat niitä esimerkiksi kirjastokokoelmista. Esimerkiksi nimeke- ja tekijätietojen merkitys hakumetatietona sekä automaattisen sisällönkuvailun mahdollisuudet ovat tutkimusaineistoilla usein erilaisia kuin vaikkapa tutkimusartikkeleilla ja monografioilla. Sähköisten tutkimusaineistojen arkistointiin on myös oma metatietostandardi *Data Documentation Initiative, DDI* (DDI 2015).

Tutkimusaineistojen sähköinen arkistointi on myös aihetta kuvailevan metatiedon ja siihen liittyvien hakutermien suhteen kiinnostava tutkimuskohde. Leimallisena piirteenä Yhteiskuntatieteellisen tietoarkiston aineistoissa on muun muassa se, että aiheenmukaisen haun mahdollistaa yksinomaan arkiston tarjoama, arkiston työntekijöiden tuottama, metatieto: asiasanoitukset, tiivistelmät ja luokitukset. Tekijöiden tuottamaa avainsanoitusta, käyttäjien tuottamaa kuvailua (folksonomiaa) tai automaattista sisällönkuvailua ei ole käytössä. Nämä aiheita kuvailevan metatiedon vaihtoehdot ovat saaneet kirjastomaailmassa enemmän jalansijaa kuin arkistoissa.

Toisaalta Yhteiskuntatieteellinen tietoarkisto on esimerkki sähköisestä aineistoarkistosta, joka noudattaa aineistojensa kuvailussa osin samoja periaatteita ja standardeja kuin kirjastot. Sen aineistokuvailuissa käytetään esimerkiksi Kansalliskirjaston ylläpitämää Yleistä suomalaista asiasanastoa YSAa, johon Suomessa sekä yleiset että tieteelliset kirjastot perustavat asiasanoituksensa.

Kirjastomaailmassa itsepalvelu on tullut vallitsevaksi kokoelmista tehtävän tiedonhaun tavaksi jo arkistoja aiemmin. Niinpä kirjastoissa saadut kokemukset asiakkaiden hakukäyttäytymisestä ja informaatioalan ammattilaisille suunniteltujen metatiedon työkalujen, kuten asiasanastojen, toimivuudesta asiakkaiden omatoimisessa haussa voivat olla arkistoille varsin huomionarvoisia ja hyödyllisiä. Kirjastojen asiakkaiden tekemiä hakuja tarkastelleet lokianalyysitutkimukset ovatkin tässä työssä olennainen taustalukemisto.

Luvussa 2 esittelen Yhteiskuntatieteellistä tietoarkistoa ja sen *Aila* -aineistoportaalia sekä esitän tutkimuskysymykset. Työn kolmas luku käsittelee aiempaa tutkimusta ja tutkimukseni keskeisiä

käsitteitä. Sen lopuksi esitän aiemman tutkimuksen pohjalta laatimani tutkimushypoteesit. Neljännessä luvussa esittelen tutkimusaineiston ja sen käsittelyssä käyttämäni menetelmät. Viidennessä luvussa esittelen tulokset. Luvussa kerrotaan ensinnäkin eri hakutapojen osuudet ja toiseksi analysoidaan tarkemmin aiheenmukaisia hakuja eli asiasana- ja vapaasanahakuja. Kuudennessa luvussa pohdin tulosten merkitystä ja suhdetta aiempaan tutkimukseen. Esitän lopuksi, luvussa 7, myös huomioita sekä hakujärjestelmän muutostarpeista että mahdollisen tulevan tutkimuksen aiheista.

2 AIHE JA TUTKIMUSKYSYMYKSET: YHTEISKUNTATIETEELLINEN TIETOARKISTO JA AINEISTOHAUT TIETOJÄRJESTELMÄSTÄ

Tässä luvussa esitellään tutkimuksen aihe ja konteksti. Luvussa 2.1. kerrotaan Yhteiskuntatieteellisestä tietoarkistosta (FSD), jonka aineistoportaali *Ailasta* tutkimuksessa käytetty lokiaineisto on saatu. Alaluvussa 2.2. esitellään *Aila*-aineistoportaali ja alaluvussa 2.3. kerrotaan FSD:n tuottamasta aiheita kuvailevasta metatiedosta, johon alaluvussa 2.4. esitellyt tutkimuskysymykset osin liittyvät.

2.1 Yhteiskuntatieteellinen tietoarkisto

Yhteiskuntatieteellinen tietoarkisto (*Finnish Social Data Archive, FSD*) tallentaa sähköisiä tutkimusaineistoja ja tarjoaa niitä jatkokäyttöön. Vuonna 1999 perustettu arkisto toimii Tampereen yliopiston yhteydessä opetus- ja kulttuuriministeriön rahoittamana ja sen palvelut ovat maksuttomia. Yhteiskuntatieteiden lisäksi arkisto on viime vuosina alkanut tallentaa yhä enemmän myös muiden ihmistieteellisten alojen aineistoja. Säilytettäväksi siis tulee entistä enemmän ja sisällöltään entistä moninaisempia uusia aineistoja. Toisaalta poistoja ei juuri tehdä. Arkistonmuodostussuunnitelmassa niistä ei ole mitään mainintaa. Aineistomääristä kerrotaan, että sähköisten tutkimusaineistojen ja niihin liittyvän muun sähköisen materiaalin kertymä vuodesta 1999 huhtikuuhun 2014 on 45 gigatavua. Kertymäennuste on noin 1 gigatavu vuodessa (Yhteiskuntatieteellisen tietoarkiston arkistonmuodostussuunnitelma).

Tutkimusaineistojen sähköisen arkistoinnin edelläkävijänä Suomessa FSD on toteuttanut kahtalaista tehtävää: ensinnäkin se on aktiivisesti pyrkinyt saamaan tutkijoita tallettamaan aineistonsa arkistoon ja toiseksi aktiivisesti tarjonnut niitä uudelleenkäyttöön. Tietoarkisto on mukana myös *Finna*-hankkeessa, jossa kootaan arkistojen, kirjastojen ja museoiden aineistoja yhteisen palvelun kautta haettaviksi. Tässä vaiheessa tietoarkiston aineistot eivät kuitenkaan ole haettavissa Finnan kautta.

FSD:n aineistoluettelon aiemmassa, kevääseen 2014 saakka käytössä olleessa asiakaskäyttöliittymässä asiakkaat pääsivät tarkastelemaan arkiston tarjontaa, mutta aineiston käyttöön saaminen edellytti yhteydenottoa arkistoon. Vanhan asiakaskäyttöliittymän korvannut palveluportaali *Aila* otettiin käyttöön toukokuussa 2014. Rekisteröityneet asiakkaat voivat sen kautta ladata aineistoja suoraan käyttöönsä. Lisäksi mukana on joitakin avoimia aineistoja, jotka voi ladata myös rekisteröitymättä.

Tietyt aineistot ovat olleet avoimesti käytettävissä jo aiemminkin ja arkisto on seurannut niiden käyttömääriä palvelinlokista (Yhteiskuntatieteellinen tietoarkisto 2014).

Asiakkaat voivat siis *Ailan* myötä hakea ja saada käyttöönsä aineistoja ilman yhteydenottoa arkiston henkilökuntaan, joka neuvoisi niiden löytämisessä. Tässä vaiheessa metatiedon ja toimivien hakumahdollisuuksien merkitys korostuu. Tarkan sisällönkuvailun avulla voidaan osaltaan pitää hakujen tuottamat tulosjoukot kooltaan kohtuullisina ja näin vaalia aineistojen tosiasiallista löydettävyyttä.

Huomionarvoista on, että tällä hetkellä *Ailan* haussa ei ole mahdollisuutta tehdä ensimmäiseen tulosjoukkoon lisärajauksia. Haun tarkkuudella ja järjestelmän relevanssilajittelulla on siten monessa tapauksessa suuri vaikutus tosiasialliseen käytettävyyteen.

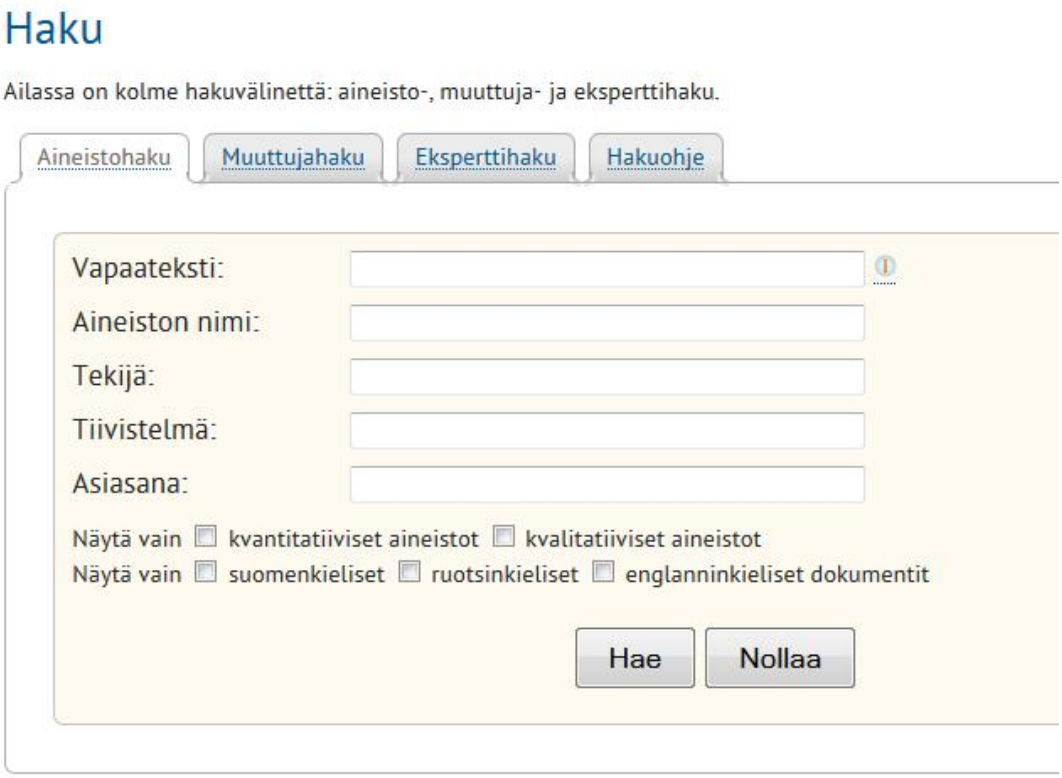
2.2 Aila-aineistoportaali

FSD:n aiemman asiakaskäyttöliittymän korvannut palveluportaali *Aila* otettiin käyttöön keväällä 2014. Rekisteröityneet asiakkaat voivat sen kautta ladata aineistoja suoraan käyttöönsä. Lisäksi mukana on joitakin avoimia aineistoja, jotka voi ladata myös rekisteröitymättä. *Aila* sai heti alkuun paljon käyttäjiä, ja tammikuussa 2015 FSD kertoi sen lisänneen aineistojen käyttöä ja asiakasmääriä huomattavasti (Yhteiskuntatieteellinen tietoarkisto 2015).

Aiemmassa aineistoluettelon käyttöliittymässä asiakkaat pääsivät tarkastelemaan arkiston tarjontaa, mutta aineiston käyttöön saaminen edellytti yhteydenottoa arkistoon. *Ailan* myötä asiakkaat voivat hakea ja saada käyttöönsä aineistoja ilman yhteydenottoa arkiston henkilökuntaan, joka neuvoisi niiden löytämisessä. *Ailan* myötä siis korostuu hakua palvelevan metatiedon (*discovery metadata)* ja asiakkaiden tiedonhakutaitojen vaikutus aineistojen löytymiseen.

Sekä itsepalvelu että arkiston jatkuvasti karttuva aineistokokoelma edellyttävät laadukasta ja riittävän tarkkaa kuvailevaa metatietoa onnistuneen haun mahdollistamiseksi. Sen avulla voidaan osaltaan pitää saannit eli hakujen tuottamat tulosjoukot kooltaan kohtuullisina ja näin vaalia aineistojen tosiasiallista löydettävyyttä.

Ailassa voi sekä selailla kaikki aineistot kattavaa aineistoluetteloa että tehdä erilaisia komentohakuja. Luettelossa oli lokakuussa 2014 aineistoja 1092 kpl eli 11 sivua. Luettelon voi järjestää aineiston numeron, nimen, tyypin (kvalitatiivinen tai kvantitatiivinen) tai julkaisupäivämäärän mukaan.



KUVA 1: *Ailan* yleishaun (*Aineistohaku)* hakulomake.

Ailan lisäksi pääsee FSD:n verkkosivujen kautta hakemaan aineistoja myös *Nesstar*-aineistotietokannasta, jossa haun voi kohdentaa esimerkiksi otsikkoon, asiasanoihin, muuttujien nimiin tai kysymysteksteihin. Tätä hakumahdollisuutta ovat asiakkaat käyttäneet melko vähän ja halutessaan ladata aineistoja käyttöönsä on heidän joka tapauksessa haettava ne *Ailan* kautta. Tässä tutkimuksessani keskityn kokonaan *Ailaan* ja jätän Nesstarin huomiotta.

2.3 Aiheita kuvaileva metatieto tietoarkiston luettelossa

Tietoarkiston aineistojen sisällönkuvailussa käytetään FSD:llä vapaamuotoisia kuvailutekstejä (aineistokuvailu), asiasanoitusta ja luokitusta. Luokituksen mukaan ei voi hakea, vaan yksinomaan selailla.

Asiasanoituksessa käytetään suomeksi Yleistä suomalaista asiasanastoa YSAa ja englanniksi yhteiskuntatieteiden yhteiseurooppalaista ELLST -tesaurusta. Luokittelussa käytetään FSD:n omaa tieteenalaluokitusta ja data-arkistojen yhteiseurooppalaista CESSDAn aihepiiriluokitusta.

Asiasanoitus on melko runsasta: asiasanoja on usein yli 10 ja samaa aihepiiriä kuvaamaan on valittu useampikin termi, kuten kuvan esimerkissä *isyys* ja *isät* sekä *sota*, *sota-aika* ja *sotatoimet (*sotatoimista aineisto ei kuvauksen perusteella kerro).

KUVA 2: esimerkki aineistokuvailusta. Näkyvissä tiivistelmä (*Sisällön kuvaus*) ja asiasanoitus.

Kuvan ulkopuolelle jää muu metatieto. Kustakin aineistosta kerrotaan kuvailussa aiheen lisäksi myös muun muassa aineiston tyyppiin, kerääjään, keruuajankohtaan, käyttöoikeuksiin, laajuuteen ja tiedostomuotoihin liittyvää tietoa.

2.3.1 Aiheita kuvaileva metatieto haun tukena tietojärjestelmässä

Tässä tutkimuksessa ei käsitellä luokitukseen perustuvaa selailuhakua ja siihen liittyvä kuvaileva metatieto, luokitus, haun tukena jää tarkastelun ulkopuolelle. Tämän työn kannalta merkittävin aihetta kuvailevan metatiedon osa-alue on asiasanoitus.

Ailan asiasanahaku kohdistuu asiasanakenttään ja tuottaa osumia asiasanakuvailuun, josta esimerkki näkyy kuvassa 2. Hakuohjeissa kerrotaan, että käytössä ovat YSA- ja ELLST -asiasanastot. Teksti ”Käytetyt asiasanastot: Yleinen suomalainen asiasanasto (YSA), ELSST-tesaurus, FSD:n tieteenala-luokitus, CESSDAn aihepiiriluokitus” (ks. liite 1) johtaa harhaan sikäli, että siinä kutsutaan myös luokituksia asiasanastoiksi.

Käyttäjää ei *Ailan* hakuohjeessa ohjata linkeillä asiasanastoihin. Tiedonhakujärjestelmä ei myös-kään ohjaa asiasanastoon siten, että se ehdottaisi asiasanahaussa hakutermeinä käytettyjen ei-asiasanojen tilalle asiasanoja.

Näin ollen voidaan todeta, että tämän hetkinen tiedonhakujärjestelmä ei juuri tue aiheita kuvailevan metatiedon käyttöä *Ailan* haussa asiasanoituksen osalta. Vapaasanahaun tuloksellisuus puolestaan perustuu pitkälti *sisällön kuvaus* -tiivistelmiin, joissa esiintyviin sanoihin saadaan osumia va-paasanahaussa. Tiivistelmät ovatkin tämän hetkisessä *Ailassa* haun kannalta merkittävin aiheita kuvailevan metatiedon osa-alue.

2.4 Tutkimuskysymykset

A-ryhmän kysymyksiin etsin vastausta kvantitatiivisin menetelmin. B-ryhmän kysymyksiin vastaan satunnaisotannan, esimerkkipoimintojen ja omien arvioideni pohjalta pyrkimättä esittämään katta-via päätelmiä. B-ryhmän kysymyksiin paneudun lähinnä aineiston laadullisen tarkastelun keinoin.

A Millaisia hakuja *Ailan* käyttäjät tekevät?

Miten paljon asiakkaat käyttävät eri hakulomakkeita eli ekspertti-, muuttuja- ja yleishakua? Kuinka suuri on asiasanahakujen ja toisaalta vapaasanahakujen osuus yleishauista? Entä kuinka suuri osa asiasanahaussa käytetyistä termeistä on asiasanaston asiasanoja?

B Hakuominaisuudet, aihetta kuvaileva metatieto ja löydettävyys lokin valossa

Mitä viitteitä lokitiedot antavat aineistojen aiheita kuvailevan metatiedon vaikutuksista niiden löydettävyyteen? Mitä vapaasana- ja asiasanahakujen laadullinen tarkastelu kertoo tietojärjestelmän hakuominaisuuksien ja osaltaan myös käytettyjen dokumentaatiokielten toimivuudesta?

3 KÄSITTEET, AIEMPI TUTKIMUS JA HYPOTEESIT

Tässä luvussa esitellään työn keskeiset käsitteet ja katsaus aiempaan tutkimukseen. Aliluku 3.1. käsittelee käsitteitä ja aliluku 3.2 aiempaa tutkimusta. Lisäksi siinä esitellään lyhyesti lokianalyysimenetelmän ominaispiirteitä sekä käyttömahdollisuuksia tiedon organisoinnin ja tiedonhakukäyttäytymisen tutkimuksessa. Aliluvussa 3.3. esittelen tämän työn tutkimushypoteeseiksi poimimiani aiempien tutkimusten havaintoja.

3.1 Käsitteet

Metatieto on määrämuotoista ”tietoa tiedosta”. Tässä työssä metatiedon käsite merkitsee Hiderin määrittelemään tapaan samaa kuin *informaatioresurssien kuvailu* (Hider 2012, 4). Metatiedolla on monia käyttötarkoituksia, joista on informaatiotutkimuksessa kehitelty erilaisia malleja. Tässä hyödynnän Haynesin viiden kohdan mallia, jonka mukaan metatietoa tarvitaan

- informaatioresurssien kuvailuun
- informaatioresurssien hakuun
- informaatioresurssien hallinnointiin
- informaatioresurssien omistusoikeuden ja aitouden määrittämiseen
- informaatioresurssien tietojärjestelmällisen yhteensopivuuden helpottamiseen

(Haynes 2004, 15–16, vrt. Chowdury & Chowdury 2007, 141. Chowdury & Chowdury kiinnittävät huomiota myös metatiedon merkitykseen käyttöoikeuden määrittämisessä ja käyttöhistorian dokumentoinnissa, resurssien sisällön pysyvyyden varmistamisen tekijänä sekä rakenteen ja asiayhteyden määrittämisessä. Nämä metatiedon osa-alueet ovat nekin huomionarvoisia sähköisten tutkimusaineistojen arkistoinnissa).

Tässä työssä tarkastellaan metatietoa *tiedonhaun (information retrieval)* näkökulmasta. Haynesin malli on tutkimusasetelman kannalta huomionarvoinen erityisesti siksi, että siinä on määritelty *kuvailu* ja haku toisistaan erilliseksi. Haynesin mallista nähdään, että kuvailulla on myös muita tarkoituksia kuin haun mahdollistaminen. Tässä työssä tarkastelen informaatioresurssien *sisällönkuvailua* yksinomaan haun näkökulmasta. Tutkimukseni kannalta olennainen metatiedon puoli on ”hakumetatieto” (*discovery metadata,* Haynes 2004, 13) eli *hakuelementteinä* toimiva metatieto. Hakuelementit ovat tyypillisesti standardisoituja, kun taas (muussa) kuvailussa yleensä noudatetaan dokumentille uskollisena pysymisen periaatetta. Hakuelementteinä käytetään ilmauksia, jotka eivät vält-

tämättä löydy suoraan dokumentista, vaan esimerkiksi kontrolloidusta asiasanastosta. Hakuelementtejä siis luodaan kääntämällä dokumentista löydettäviä ilmauksia ja/tai aiheita dokumentaatiokielelle. Dokumentille uskollisessa kuvailussa taas säilytetään esimerkiksi kirjoitusvirheet ja standardeista poikkeavat ei-länsimaisten kirjoitusasujen translitteroinnit (Suominen et al 2009, 45 – 46).

> The field of information organization deals with documentation packaged up as resources, which are also commonly referred to as *documents*, even when they contain non-textual information. Information organization is thus concerned with *document retrieval.*
>
> (Hider 2012, 13)

Dokumentti toimii tiedon organisoinnin yksikkönä. Hyvin monenlaiset informaatioresurssit voidaan tiedon organisoinnin alalla lukea dokumenteiksi: teksti tai kuva, fyysinen tai sähköinen tallenne, ohjelmisto, luonnonobjekti tai esine ja niin edelleen (Suominen 2009, 21–25). Niinpä myös tässä tutkimuksessa tarkasteltavan tiedonhaun kohteet, sähköiset tutkimusaineistot, ovat dokumentteja.

”Tyypillinen, mutta ei kuitenkaan ainoa sisällöllinen piirre, jonka perusteella dokumentteja haetaan, on dokumentin aihe eli se, mistä dokumentti 'kertoo'” (Suominen et al 2009, 137). Dokumenttien sisällönkuvailu on siis osa metatietoa. Aiheen mukaisen haun kannalta tarpeellista, aihetta kuvaavaa metatietoa (*subject metadata,* Zavalina 2011, 104) ovat asiasanoitukset, tiivistelmät ja luokitukset. Asiasanoitus perustuu kontrolloituihin sanastoihin, *asiasanastoihin.* Asiasanastoja ja luokituksia käytetään sisällönkuvailussa *dokumentaatiokielinä* eli *indeksointikielinä* (Suominen et al 2009, 190; Vakkari 1999, 22).

YSA- ja ELLST -asiasanastot ovat *tesauruksia.* Tesaurukset perustuvat sanojen ryhmittelyyn ja keskinäisten suhteiden määrittelyyn. Tiedonhakutesaurusten standardin (ISO 25964-1) mukaan tesauruksessa tulee muun muassa määritellä termien ekvivalenssi- hierarkia- ja assosiatiiviset suhteet. Synonyymeja ja lähikäsitteitä kontrolloidaan *ohjaustermeillä (lead-in term).*

KUVA 3: Esimerkki termien suhteista YSA -tesauruksessa.

Kuvassa oikealla näkyy assosiatiivinen suhde *rinnakkaistermit* ja hierarkkinen suhde *laajemmat termit*. Ohjaustermillä puolestaan ohjataan käyttämään tiettyä yhtä sanaa käsitteistä, joiden välillä on ekvivalenssisuhde eli jotka tarkoittavat luonnollisessa kielessä samaa tai lähes samaa. Tästä esimerkki vasemmalla: ”a priori tieto käytä: apriorinen tieto”.

Luonnollisen kielen sanoja käytetään tai voidaan asiasanastoissa käyttää muusta kielen käytöstä eroavin tavoin. Kuten Hider toteaa, voi standardoitu sanasto parantaa *tiedonhaun* tuloksia – etenkin jos sekä kuvailija että hakija käyttävät samaa sanastoa. Siinä, missä tiettyä ilmiötä voi kuvata monilla luonnollisen kielen sanoilla, voidaan ihannetapauksessa nämä kaikki löytää haussa yhdellä dokumentaatiokielen sanalla. ”*The aim is for each concept to be represented by one, and only one, particular term, and for each term to mean only one particular concept*” (Hider 2012, 151).

Samassa yhteydessä Hider kuitenkin huomioi kuvailun rajoitteet muistuttaen, että aihe ei aina ole helposti määriteltävissä: ”*there are different ways in which subject analysis can be translated into the terms of a controlled vocabulary. For example, a resource may cover multiple topics, but usually only some of them will be represented*” (Hider 2012, 151). Tämä liittyy *aboutnessin* määrittämiseen eli dokumentaatiokielen käyttöön liittyvään aiheen pelkistämiseen – sen ratkaisemiseen, mitä dokumentista löytyvistä aiheista pidetään ”ydinsanomana” ja sisällytetään kuvailevaan metatietoon.

> ...yksittäisen dokumentin sisältöön liittyvien asioiden moninaisuuskin on yleensä varsin huomattava verrattuna siihen, mitä dokumentin sisällönkuvailussa voidaan tuoda siitä esiin. Tätä dokumentaatiokielen ja sisällönkuvailun pelkistävää suhdetta on alan kirjallisuudessa kuvattu käsitteellä *aboutness.* (Suominen et al 2009, 142)

Avainsanalla (keyword) voidaan tarkoittaa suoraan tekstistä poimittua luonnollisen kielen sanaa, jota perusmuotoistettuna voidaan käyttää myös tekstin sisällön kuvailuun. *Avainsanahausta* kuitenkin puhutaan usein myös ylipäätään luonnollisen kielen sanoja käyttävän *vapaasanahaun (free-text search)* synonyymina (ks. esim. Suominen et al 2009, 343–345). Tässä työssä tarkoitan avainsana- ja vapaasanahaulla samaa asiaa.

Tiedonhaulla tarkoitetaan tässä tutkimuksessa hakutehtävän suorituksessa syntyvää *hakuprosessia* (Järvelin & Sormunen 2006, 110). Käyttäjien *tiedonhakujärjestelmässä* tekemät *kyselyt (query)* ovat *hakusuorituksia (retrieval performance,* Haynes 2004, 17*)*. Tutkimukseni tuottaman tiedon voidaan katsoa olevan juuri tietoa hakusuorituksista. Käyttäjien tekemät kyselyt hakusuorituksina ovat osa *tiedonhakukäyttäytymistä*, joka kuitenkin on laajempi käsite. En siis tässä esitä tutkivani asiakkaiden tiedonhakukäyttäytymistä kokonaisuutena. FSD:n tietojärjestelmässä tehdystä tiedonhausta ovat aineistossani mukana komentohaut. Selailuhakua, joka myös on järjestelmässä mahdollista, en tässä tutki. En myöskään jäljitä yksittäisten käyttäjien *sessioita* eli tutkimukseni ei tuota tietoa siitä, mitkä aineiston hauista ovat yksittäisen käyttäjän saman istunnon aikana tekemiä. Tiedonhakukäyttäytymisen tutkiminen kokonaisuutena edellyttäisi vähintäänkin molempien hakutapojen, selailu- ja komentohaun, sekä käyttäjäkohtaisten hakusessioiden tarkastelua.

Tämän tutkimuksen aineisto on koostettu tiedonhakujärjestelmän lokitiedoista *(transaction logs)*. Lokitiedot ovat sähköisesti tallennettua jälkeä tietojärjestelmän ja sen käyttäjien välisestä kommunikaatiosta (Jansen 2006, 408.) Tässä tutkimuksessa käytetään tiedonhakujärjestelmän *palvelinlokista* (*server-side log)* koostettua aineistoa, jossa on mukana hakuihin liittyvä informaatio. Palvelinlokiin tallentuva käyttäjiin liittyvä informaatio, kuten IP -tunnisteet, ei ole aineistossa mukana.

3.2 Aiempi tutkimus

Kirjastojen avointen aineistoluetteloiden (*Open Access Catalogue, OPAC)* lokien avulla on tutkittu käyttäjien tiedonhakukäyttäytymistä sekä arvioitu hakujen onnistumista ja tietojärjestelmien toimivuutta. Tutkimuksissa on selvitetty muun muassa hakujen ominaispiirteitä, kuten eri hakukenttien käytön yleisyyttä, sekä tulosmääriä. Usein on huomiota kiinnitetty nollatuloksiin eli hakuihin, jotka eivät ole johtaneet yhdenkään dokumentin löytymiseen. Kirjastojen aineistoluetteloiden lokeista on selvitetty muun muassa kokonaiskäyttömääriä ja käytön ajoittumista, komento- ja selailuhaun yleisyyttä, hakusessioiden piirteitä ja aihetta kuvailevan metatiedon (*subject metadata*) merkitystä hakujen onnistumisessa. Näistä viimemainittu on oman tutkimukseni kannalta olennainen. Erityisen kiinnostavaa työni kannalta olisi aineistoarkistoihin liittyvä tutkimus yleisesti ja sähköisten tutkimusaineistojen arkistoihin liittyvä tutkimus erityisesti. Ensin mainittua olen löytänyt muutaman (Pynnönen 2000, Zavalina 2011, Huurnink et al 2010), viimemainittua en ollenkaan. Aiemman tutkimuksen puuttumista sähköisten aineistoarkistojen osalta selittää muun muassa se, että ala on kirjastoihin verrattuna nuori ja pieni.

Jo ensimmäisten kirjastomaailmassa tehtyjen asiakkaiden hakukäyttäytymistä selvittäneiden tutkimusten aikoihin on alan julkaisuissa keskusteltu aihetta kuvailevan metatiedon kysymyksistä ja erityisesti asiasanotuksen tulevaisuudesta. Asiakkaiden itsenäisesti käytettäviksi asetettujen avointen aineistoluetteloiden hakuominaisuudet perustuivat, ja perustuvat pitkälti edelleen, kirjastoammattilaisten luomaan ja heidän käyttöönsä suunniteltuun metatietoon. Lokiaineistot osoittivat jo 1990 -luvulla, että asiakkaat eivät välttämättä osanneet aiheen mukaisissa hauissaan näitä hyödyntää. Larsonin vuonna 1991 julkaistuun pitkittäistutkimukseen viitataan yhä. Hänen otsikkoonsa tiivistyy tulos ja ennuste: *The Decline of Subject Searching.* Larson listasi yleisimpiä tiedonhakuongelmia:

* Käyttäjät tuntevat asiasanastot puutteellisesti.

* Käyttäjillä on vaikeuksia hakulausekkeiden muotoilun mekaanisessa ja käsitteellisessä puolessa.

* Haut eivät tuota tuloksia.

* Haut tuottavat liikaa tuloksia.

* Haut tuottavat epärelevantteja tuloksia.

(Larson 1991, viitattu Yu & Yang 2004. Vapaasti kääntämäni).

Tämä Larsonin lista on kiinnostava myös suhteessa Huuskosen ja Vakkarin tutkimukseen, jossa arvioitiin asiasanaston tuntemuksen merkitystä opintosuorituksen onnistumisessa. Koeasetelmassa lääketieteen opiskelijat etsivät aineistoa tieteellisistä lehtitietokannoista ja tekivät sen pohjalta esseitä tutkijoiden tarkastellessa prosessia aina hakujen lokeista esseiden arviointiin saakka. He totesivat, etteivät *Medical Subject Headings* -asiasanoja hauissaan käyttäneet opiskelijat saaneet muita parempia arvosanoja (Huuskonen & Vakkari 2007).

Huuskosen ja Vakkarin tutkimus on harvoja löytämiäni Suomessa tehtyjä tutkimuksia, jossa lokiaineistosta on selvitetty erityisesti hakutermeihin liittyviä kysymyksiä. Mainitsemisen arvoinen tässä yhteydessä on myös Mukalan (2005) pro gradu -tutkimus, jossa tarkasteltiin hakulausekkeiden muotoilua ja hakujen tuloksellisuutta vapaasanahakuun perustuvassa tiedonhakupelissä. Dokumentaatiokielisiä hakutermejä ei tämäkään tutkimus kuitenkaan käsitellyt. Käyttäjien kykyä hyödyntää dokumentaatiokieltä hakulausekkeiden muotoilussa ovat tutkineet Sihvonen ja Vakkari (2004), joiden tutkimuksessa vertailtiin alan eli kasvatustieteen hyvin tuntevien ja noviisien tekemiä hakuja kasvatustieteen ERIC -tietokannassa. Sihvosen ja Vakkarin mukaan asiasanaston termejä pystyivät hakulausekkeiden muotoilussa parhaiten hyödyntämään aihepiiriä ennestään tuntevat käyttäjät (Sihvonen & Vakkari 2004, 688). On huomattava, että ERIC -tietokannan asiasanahaussa järjestelmä ohjaa käyttäjiä asiasanaston termeihin. Asiasanahaun edellytykset ovat siis toisenlaiset kuin esimerkiksi *Ailassa*, jossa tällaista ominaisuutta ei ole.

Muu tutkimukseni taustakirjallisuus tulee paljolti Yhdysvalloista ja erityisesti tieteellisten kirjastojen maailmasta. Hyödynnän sitä työssäni olettaen, että osa asiakkaiden hakukäyttäytymistä ja aihetta kuvailevan metatiedon toimivuutta koskevista havainnoista on yleistettävissä ja/tai verrattavissa arkistokontekstiin. Arkistoissa myös käytetään usein samoja dokumentaatiokieliä ja muita aihetta kuvailevan metatiedon välineitä kuin kirjastoissa.

Kirjastomaailmassa on pitkään keskusteltu asiasanoituksen tulevaisuudesta. Myös lokianalyyseista on etsitty vastausta kysymyksiin erityyppisten hakumahdollisuuksien toimivuudesta asiakkaiden itsenäisesti tekemässä tiedonhaussa (ks. esim. Antell & Huang 2008, Gross & Taylor 2005).

Eräissä kirjastoasiakkaiden hakuja kartoittaneissa lokitutkimuksissa on todettu sekä vapaasana- että asiasanahakujen olleen harvinaisia. Esimerkiksi Columbian yliopiston opettajakoulutuslaitoksen kirjaston lokeja tutkineet Asunka ja muut saivat asiasanahakujen osuudeksi vain 4 prosenttia ja avainsanahakujen sitäkin vähemmän, 3 prosenttia. Heidän aineistossaan suosituimpia hakukenttiä

olivat otsikko (41 prosenttia) sekä kaikkeen metatietoon kohdistuva ”mikä tahansa”(35 prosenttia). Myös tekijähaku, 13 prosenttia, oli selvästi aiheenmukaista hakua yleisempää (Asunka et al 2009, 41–42). Samoin Antellin ja Huangin Oklahoman yliopistokirjastossa tehdyssä käyttäjätutkimuksessa asiasanahaut olivat harvinaisia: 4,6 prosenttia kaikista hauista. Asiasanahakujen onnistumisprosentiksi Antell ja Huang laskivat 54,8. Käytetyin hakukenttä heidän aineistossaan oli avainsanahaku 64,8 prosentin osuudella kaikista hauista. Seuraaviksi yleisimmät olivat noin 12 prosentin osuuksilla tekijä- ja otsikkohaut. (Antell & Huang 2008, 71).

Villén-Rueda ja muut tarkastelivat tutkimuksessaan Granadan yliopiston kirjaston lokeja selvittäen eri hakutyyppien käyttömääriä ja hakujen tuloksellisuutta eri asiakasryhmillä. Myös he kiinnittivät huomiota asiasanahaun ja muun aiheenmukaisen haun ongelmiin. Heidän tutkimuksessaan aiheenmukaisen haun osuus oli suurempi kuin edellä mainituissa, 14 prosenttia. Suosituimpia olivat kuitenkin nimekkeen ja tekijän mukaiset haut, 49 ja 37 prosenttia. Villén-Rueda ja muut myös totesivat haun aihepiiriä hyvin ennalta tuntevien asiakkaiden eli professorien erityisesti suosivan tekijä- ja otsikkohakuja (Villén-Rueda et al 2007, 335–336).

Verkkohaun tuomia hakukäyttäytymisen muutoksia kirjastokontekstissa tutkineet McKay ja Buchanan (2013, s. 498, ks.myös Lau 2006) vetävät aiempien tutkimusten tuloksia yhteen viitaten sekä hakukoneita että verkkokirjastoja käsitelleisiin töihin. Niiden pohjalta he toteavat, että käyttäjät tekevät lyhyitä ja yksinkertaisia hakulausekkeita, eivät muuta oletusasetuksia eivätkä käytä tarkennettua hakua. Kiinnostavaa on, että McKayn ja Buchananin omassa tutkimuksessa käyttäjät toimivat osaksi toisin (McKay & Buchanan 2011, 504).

McKay ja Buchanan esittävät muutenkin referoimassaan tutkimusperinteessä vallitsevia käsityksiä optimistisempia arvioita käyttäjien tiedonhakukäyttäytymisestä ja -taidoista. He kiinnittävät huomiota käyttäjien toiminnan rationaalisuuteen ja käytännöllisyyteen. He edustavat myös selkeästi näkemystä, jonka mukaan tiedonhakujärjestelmien on nyt sopeuduttava hakukoneiden muovaamiin tiedonhaun tapoihin ja korostavat avainsanahaun merkitystä. McKayn ja Buchananin työtä voi pitää esimerkkeinä aiempaa kirjastomaailmassa tehtyä käytettävyystutkimusta verkkosuuntautuneemmasta, luettelokriittisemmästä ja avainsanakeskeisemmästä tutkimusotteesta.

Jos McKay ja Buchanan todella ovat tyypillisiä hakujärjestelmien käytettävyystutkimuksen nykysuuntien edustajia, ovat käyttäjien toimintaan keskittyvien kysymysten rinnalla ja tilalla nyt vahvasti kysymykset siitä, kuinka itse tietojärjestelmien ja niiden asiakaskäyttöliittymien on muututtava

ja mukauduttava. Lokianalyysitutkimukset ovat, osaltaan, jo pitkään osoittaneet jonkinlaisen muutoksen olevan tarpeen: hakutuloksissa on parantamisen varaa (ks.esim. Larson 1991, Yu & Young 2004, Villén-Rueda et al 2007, Antell & Huang 2008).

McKay ja Buchanan kyseenalaistavat aiemmin esitetyt keinot hakutulosten parantamiseksi esimerkiksi käyttäjäkoulutusta lisäämällä. Antell ja Huang puolestaan näkevät käyttäjien opastamisen yhtenä ratkaisuna, mutta esittävät radikaaleja muutoksia opastuksen tapoihin (Antell & Huang 2008, 75). Kolmas huomionarvoinen näkökulma on taloustieteellisen asiasanaston merkitystä hauissa tutkineella Borstilla. Hän edustaa Saksan taloustieteellistä kansalliskirjastoa eli kontekstina on tiettyyn alaan keskittynyt kokoelma ja sanasto. Borst raportoi lokianalyysitutkimuksesta, jonka mukaan käyttäjät käyttävät asiasanoja melko usein ja hyötyvät niistä – mikäli sanasto on laaja ja hyvin päivitetty. Hän katsoo asiasanoituksen olevan tulevaisuudessakin tarpeellista ja sen käyttöä voitavan helpottaa teknisillä ratkaisuilla käyttöliittymissä (Borst 2012, 451–452).

Asiasanoituksen merkityksestä hakujen onnistumisessa ovat kirjoittaneet Gross ja Taylor (2005), jotka ovat tutkineet vapaasanahakuja ja niiden tuloksia. Yliopistokirjaston avoimen aineistoluettelon lokia analysoidessaan he huomasivat monien vapaasanahakujen tuottavan osuman asiasanoitukseen ja vain asiasanoitukseen. Tutkimuksensa pohjalta he esittävät asiasanoituksen parantavan vapaasanakentässä tehtyjen hakujen tuloksia merkittävästi: ilman asiasanoitusta jopa kolmannes heidän tutkimiensa hakujen tuloksista olisi jäänyt löytymättä.

Zavalina (2011) on tutkinut eri metatietokenttien merkittävyyttä hauissa vertaillen kolmen yhdysvaltalaisen kulttuuriperintökokoelman käyttöliittymien lokeja. Hänen mukaansa sekä avainsana- että asiasanahaut ovat asiakkaille tärkeitä heidän etsimänsä tiedon löytämisessä. Zavalinan tutkimus lähestyy arkistomaailmaa: kirjastoihin verrattuna myös kulttuuriperintökokoelmissa merkinnee nimekkeiden ja otsikoiden vähäisyys tai puuttuminen aiheenmukaisten hakujen merkityksen korostumista. Arkistomaailma onkin kiinnostava konteksti aiheenmukaisten hakujen tutkimukselle ja aihetta kuvailevan metatiedon merkityksen pohdinnalle. Olen kuitenkin toistaiseksi löytänyt yllättävän vähän arkistojen käyttöliittymissä tehtyihin lokianalyyseihin perustuvaa tutkimusta. Ainoat löytämäni tutkimusartikkelit perustuvat samaan alankomaalaisessa AV-arkistossa tehtyyn tutkimukseen, jossa on selvitetty kuvahakumahdollisuuksien parantamisen vaikutusta hakuihin (Huurnink et al 2010 ja 2011). Haun onnistumista arvioitiin osaksi sen perusteella, oliko asiakas hakunsa yhteydessä tehnyt aineistopyynnön arkistolle.

Pynnönen tutki vuonna 2000 valmistuneessa informaatiotutkimuksen gradussaan Aamulehden internet-arkistosta tehtyjä hakuja. Lokianalyysin lisäksi hän haastatteli eri käyttäjäryhmien edustajia. Tuolloinen Aamulehden internet-arkisto perustui täysin vapaasanahaulle eli asiasanoitusta ei ollut. Tämä vaikutti olennaisesti hakutuloksiin romahduttaen tarkkuuden. Tutkituista hauista suuri osa johti hyvin suureen tulosjoukkoon. Tämä aiheutti paljon haittaa, koska internet-arkistossa ei ollut myöskään relevanssilajittelua. Pynnösen työ kuvaa kiinnostavalla tavalla sähköisten arkistojen varhaisemman vaiheen käyttäjäkokemuksia ja tarjoaa osaltaan vertailupohjaa myös tähän tutkimukseen, vaikka käsitteleekin jo historiaan jäänyttä tiedonhakujärjestelmää.

Lopuksi mainittakoon myös Juha Riipisen gradu *Yhteiskuntatieteellisen tietoarkiston verkkosivujen käytettävyys noviisikäyttäjien näkökulmasta* (Riipinen 2014). Siinä on käytettävyystestauksen keinoin selvitetty FSD:n verkkosivujen käytettävyyttä, aiemman hakujärjestelmän hakuominaisuudet mukaan lukien. Tutkimus koskee *Ailan* tilalta väistyneitä hakuominaisuuksia eikä siten suoraan juuri tarjoa pohjaa tälle tutkimukselle. Jos *Ailasta* tulevaisuudessa tehdään käytettävyystestausta, on sitä varmasti mielenkiintoista verrata myös Riipisen työhön.

3.1.2 Lokianalyysi menetelmänä

Informaatiotutkijat Jansen ja Spink ovat tehneet paljon lokianalyysitutkimusta keskittyen erityisesti avoimen verkon hakukoneisiin. Heihin viitataan paljon ja he ovat julkaisseet muun muassa kattavan menetelmäoppaan lokianalyysiin (Jansen, Spink ja Taksai 2009).

Jansen (2006, 411, vapaasti kääntämäni ja kommentoimani)

vetää yhteen lokianalyysimenetelmiin kohdistettua kritiikkiä	**ja vastaa siihen**
Käyttäjän tiedontarpeet ja hänen havaintonsa ja arvionsa haun tuloksista jäävät tuntemattomiksi. Käyttäjistä ei liioin saada (demografisia) tietoja.	Näissä asioissa voidaan lokianalyysia täydentää muilla keinoin hankitulla tutkimustiedolla. (vrt. McKay & Buchanan: lokianalyysi ainoana menetelmänä on myös paljon käytetty ja puoltaa sekin paikkaansa. McKay & Buchanan 2013.)
Mielekäs toteutustapa voi olla vaikea löytää: aineiston keruun, säilyttämisen ja analyysin hankaluudet korostuvat laajoilla ja sekalaisilla aineistoilla. Tutkimuskäsitteistö ei ole ollut yhdenmukaista.	Nämä ongelmat koskevat monia muitakin menetelmiä. Nykyisin on myös entistä parempia ohjelmistoja, standardeja ja analyysimenetelmiä.
Aineisto jää vaillinaiseksi: lokitietoja saadaan ensisijaisesti palvelimilta ja selaintason toiminnot jäävät tallentumatta. Tätä ongelmaa on yritetty ratkoa esimerkiksi yhdistämällä lokianalyysiin käyttäjäkyselyjä. Käyttäjiä ei voida tunnistaa, vaikka IP-osoitteetkin olisivat käytettävissä: sama osoite voi olla monen henkilön käytössä.	Käyttäjäkyselyjen yhdistäminen lokianalyysiin kumoaa kuitenkin sen olennaisen vahvuuden: menetelmä on käyttäjälle huomaamaton eikä vaikuta hänen toimiinsa.

Selaintason mukaan saamiseksi on myös kehitetty ohjelmistoja (*Tracker*, kaupallisia *Spyware*-ohjelmistoja). Näillä voi kartoittaa yksittäisen käyttäjän hakuja käyttäjäpuolen lokista: tästä esimerkkinä Kumpulaisen tutkimus tutkijoiden tiedonhankinnasta (Kumpulainen 2013, 28, 32).

Lokianalyysitutkimuksissa on keskitytty erityyppisiin hakujen piirteisiin tutkimuskysymyksistä riippuen – näistä osan olen jättänyt tässä katsauksessa huomiotta. Lokeista voidaan tutkia istuntojen pituutta ja tältä pohjalta jaotella hakuja sekä tehdä arvioita käyttäjien päämääristä: esimerkiksi hyvin lyhyt istunto voidaan tulkita "*hit and run*" -hauksi, jossa käyttäjä jo etukäteen tietää mitä hakee. Hakulausekkeet voidaan jaotella esimerkiksi niiden sisältämien ternien määrän mukaan eri tyyppeihin tai laskea eri termien yleisyys koko aineistossa (esim. Wolfram et al 2009, 906–907).

Usein lokianalyysille on ominaista aineiston suuri tai suorastaan valtava määrä. Aineistojen käsittelemiseksi kehitetyt menetelmät ovatkin olennaisessa osassa alan kirjallisuudessa (ks. esim. Jansen 2006). Suurilla aineistomäärillä on käytetty esimerkiksi matemaattisia klusterointimalleja (esim. Wolfram et al 2009 ja Niu & Hemminger 2010). Tällaiset suurille aineistoille tarvittavat menetelmät on tässä jätetty vähäiselle huomiolle, koska tämän työn tutkimuskysymykset ja aineiston koko eivät edellytä niiden käyttöä.

3.3 Tutkimushypoteesit

Analysoituani aineiston ja vastattuani tutkimuskysymyksiini vertaan havaintojani muutamiin aiemmasta tutkimuksesta poimimiini havaintoihin. Tiivistän esittelemäni aikaisemman tutkimuksen seuraaviksi hypoteeseiksi, joita tarkastelen luvussa 6.

Hypoteesi eri hakuvaihtoehtojen käytöstä: yleisintä on yksinkertainen vapaasanahaku

Aiempaan tutkimukseen, erityisesti McKayn ja Buchananin työhön, perustuen eri hakumahdollisuuksien käyttöön liittyvänä hypoteesina on, että asiakkaat käyttävät lyhyitä ja yksinkertaisia hakulausekkeita (vrt. McKay & Buchanan 2013, 504 ja Niu & Hemminger 2010) ja suurin osa hausta on vapaasanahakua. ”Asiakkaat tapaavat käyttää yksinkertaista vasaraa tarjotun monipuolisen työkalupakin sijaan” eli valitsevat käytettävissä olevista hakuominaisuuksista vain yksinkertaisimmat (Yu & Young 2004, 174).

Hypoteesi aihetta kuvaavan metatiedon käytettävyydestä: asiasanahaku tuottaa vaikeuksia

Tutkimukseen tutustuessani havaitsin eräiden tutkijoiden raportoivan Larsonin 1990-luvulla määrittelemien asiasanahaun ongelmien kanssa samansuuntaisia havaintoja yhä 2000 -luvulla (Larson 1991, vrt. Antell & Huang 2008, MacKay & Buchanan 2013, Yu & Young 2004). Myös oma informaatioalan työkokemukseni yhtäältä aihetta kuvailevan metatiedon ja toisaalta tietopalvelun ja tiedonhaun opastuksen parissa oli saanut minut arvelemaan ”Larsonin listan” mahdollisesti kestäneen aikaa ja siinä kuvattujen tiedonhakuongelmien näkyvän nykypäivänkin lokeissa. Larsoniin ja uudempaan tutkimukseen, erityisesti Antellin ja Huangin (2008, 74) sekä Villén-Ruedan ja muiden (2007, 328) työhön pohjautuen esitän asiasanaston käyttöä koskevan hypoteesin. Hypoteesina on se, että käyttäjät tuntevat asiasanastot puutteellisesti ja heillä on vaikeuksia hakulausekkeiden muotoilun mekaanisessa ja käsitteellisessä puolessa.

Hypoteesi aihetta kuvaavan metatiedon laadun merkityksestä: lokiaineisto kertoo myös dokumentaatiokielten laadusta ja sopivuudesta

Aihetta kuvaavan metatiedon laatuun vaikuttavat sekä käytetyt välineet, tässä dokumentaatiokielet, että niiden käyttötavat. Borst on esittänyt asiakkaiden käyttävän asiasanastoja ja hyötyvän niistä, jos asiasanastot ovat riittävän laajoja ja ajantasaisia (Borst 2012).

Asiasanahaussa käytettyjen termien voi odottaa tarjoavan huomionarvoisia näkökulmia käytettyjen asiasanastojen laatuun ja kyseisen aineiston kuvailemiseen soveltuvuuteen. Tässä tutkimuksessa eivät varsinaiset tutkimuskysymykset koske asiasanastojen laatua tai soveltuvuutta. Aineisto kuitenkin tarjoaa melko laajan yleiskuvan tiettyyn tieteenalaan, yhteiskuntatieteisiin, liittyvien hakutermien käytöstä. Tätä aineistoa tutkiessani ja erityisesti asiasanahakujen ongelmakohtiin paneutuessani olen saanut ikään kuin sivutuotteena myös kuvaa asiasanastojen toimivuudesta FSD:n aineistoilla. Uskaltaudun tekemään tästä johtopäätöksiä todeten kuitenkin, että kattavamman kuvan saamiseksi olisi aineistoa tutkittava lisää ja toisenlaisin tavoin.

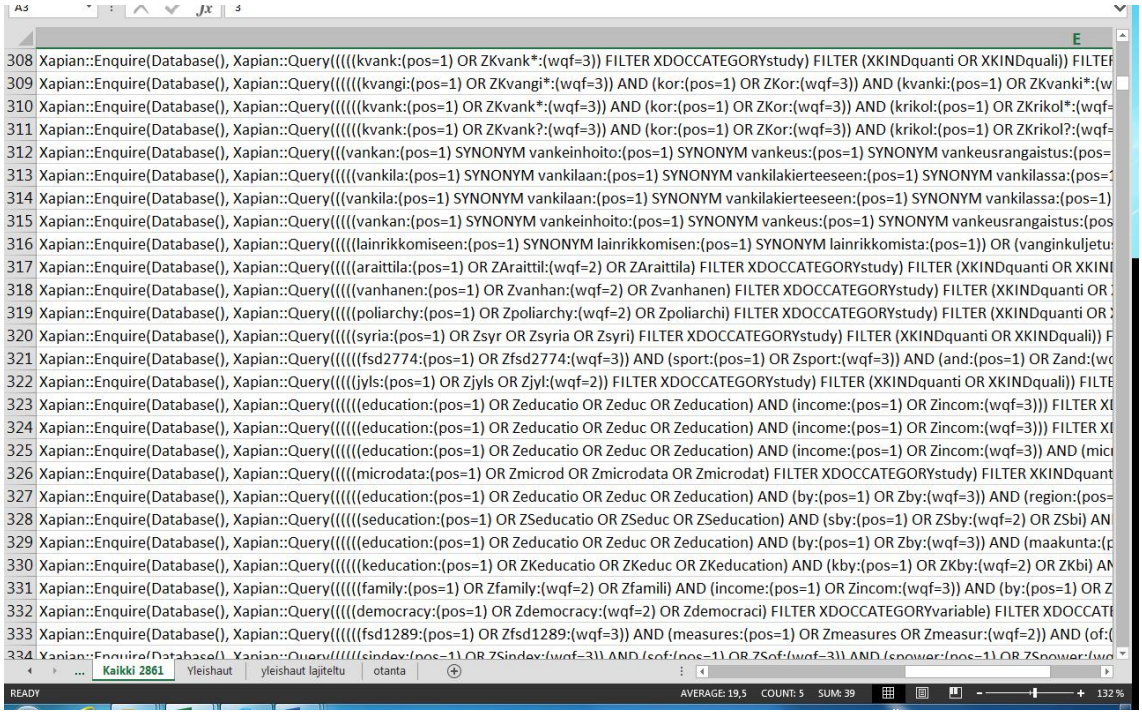
4 TUTKIMUSAINEISTO JA MENETELMÄ

Tässä luvussa esitellään tutkimusaineisto ja menetelmä. Alaluvussa 4.1. esitellään aineisto ja alaluvussa 4.2. kuvataan aineiston käsittelyä ja analyysia.

4.1 Aineisto

FSD: ltä tätä tutkimusta varten saadussa lokiaineistossa oli mukana kaikki *Ailassa* tehdyt haut sen käyttöönotosta, 12.5., alkaen 15.10. saakka. Se sisälsi 5483 hausta seuraavat tiedot

- kullekin haulle luotu tunnistenumero
- päivämäärä ja aikaleima
- käytetty hakulomake: *yleishaku (studysearch), muuttujahaku (variablesearch)* tai *ekspertti-haku (expertsearch)*
- järjestelmän generoima arvio hakutulosmäärästä. Tämä osoittautui heti ensi tarkastelussa niin paikkaansa pitämättömäksi, ettei sitä kannattanut huomioida mitenkään.
- hausta muodostetun hakulauseen tekstirepresentaatio



KUVA 4: Palvelinlokista koostettua, Excel-taulukkoon siirrettyä lokiaineistoa: näkyvissä hakulausekkeiden tekstirepresentaatioita.

Xapian viittaa hakujärjestelmän käyttämään Xapian-hakukirjastoon (http.//xapian.org). Kuvasta näkyy tapa, jolla järjestelmä muokkaa hakijoiden antamat hakutermit Boolen lausekkeiksi ja eri sijamuotoihin, joista vain osa on mukana lokista tätä tutkimusta varten koostetussa aineistossa. Kuvassa näkyvät haut 312–316 ovat esimerkkejä eksperttihausta, jossa voi kontrolloida synonyymeja. Eksperttihaku hyödyntää Xapianin *Query Parser* -jäsennintä ja edellyttää käyttäjältä Boolen logiikan tuntemusta (FSD:n atk-erikoistutkija Matti Heinonen, sähköpostitiedonanto 4.11 2014, Xapian 2015).

Kuvassa on mukana eri hakukenttiin kohdistettuja hakuja, joiden tunnisteina ovat kirjaimet

- S kohdennetaan otsikkokenttään, esimerkkinä rivi 328
- A kohdennetaan tekijäkenttään, esimerkkinä rivi 317
- K kohdennetaan asiasanakenttään, esimerkkinä rivi 330

Lisäksi on tiivistelmään, muuttujan kysymyksiin ja vastauksiin sekä kieleen kohdistettuja hakuja. Näistä ei kuvassa näy esimerkkejä.

Kaikissa hauissa mukana oleva Z merkitsee hakua stemmaamattomasta indeksistä. Haku tehdään aina sekä stemmattuun että stemmaamattomaan indeksiin, joten kunkin haun tekstirepresentaatiossa näkyy hakutermi sekä Z:lla merkittynä että ilman sitä. Lisäksi tulosdokumentteja suodatetaan filtterillä. Filtterit ovat

- XDOCCATEGORY – dokumentin tyyppi; dokumentteja ovat aineistokuvailut ja muuttujakuvailut
- XKIND – aineiston tyyppi: joko kvalitatiivinen, kvantitatiivinen tai molemmat.

Kuvasta 1 huomaa sen, että sama haku voi kirjautua lokitietoihin moneen kertaan (rivit 323 ja 324). Näin käy silloin, kun hakija palaa hakutulossivulta selaimen nuolinäppäimellä hakusivulle: tällöin haku tulee automaattisesti suoritetuksi uudestaan. Lisäksi *checkboxia* klikattaessa (*Näytä vain* -valinnat) haku suoritetaan heti uudestaan ja tuloslista päivittyy ilman *Hae* -napin painamista (Matti Heinonen, sähköpostitiedonanto 4.11.2014).

4.2 Menetelmä: määrällinen ja laadullinen analyysi

Ensin rajasin täysin samanlaisina toistuneet haut pois ja tarkasteltavaan aineistoon jäi 2861 hakua. Tutkimuksen määrälliseen osuuteen kuului hakujen jaottelu niissä käytettyjen hakulomakkeiden,

muuttuja-, ekspertti- ja yleishaun, ja yleishaun ei hakukenttien mukaisiin kategorioihin. Tämän jälkeen oli vuorossa asiasanakenttään kohdistettujen hakujen tarkastelu sekä määrällisesti että laadullisesti. Kun asiasanahaut oli erotettu muusta aineistosta, kävin läpi kaikki niissä käytetyt hakutermit selvittäen, mitkä niistä olivat FSD:n käyttämien asiasanastojen eli YSAn tai ELLSTin asiasanoja: hain jokaisen aineistossa käytetyn termin asiasanastojen verkkokäyttöliittymistä. Näin sain selville asiasanojen käytön osuuden asiasanakentän hauissa.

Hakulomakkeiden mukaisen jaottelun jälkeen keskityin yleishakuihin. Hakujen tekstirepresentaatioita voi Excel-taulukossa lajitella tunnistekirjainten avulla, mutta tulos oli varmistettava katsomalla kaikki haut läpi omin silmin. Näin sain hyvän kuvan aineistosta kokonaisuutena ja myös osviittaa hakujen suhteista toisiinsa. Tällainen aineiston käsittely jätti sijaa inhimillisille virheille, mutta toisaalta tämän kokoisella aineistolla vielä puolsi paikkaansa. Voin todeta saaneeni yleishauista, joihin huomioni erityisesti kohdistui, ja etenkin tarkimman tarkastelun saaneista asiasanahauista varsin hyvän yleiskäsityksen. Sen avulla pystyn muodostamaan aineistosta myös yleisluontoisia arvioita taulukkolaskennan avulla saadun määrällisen tiedon rinnalle.

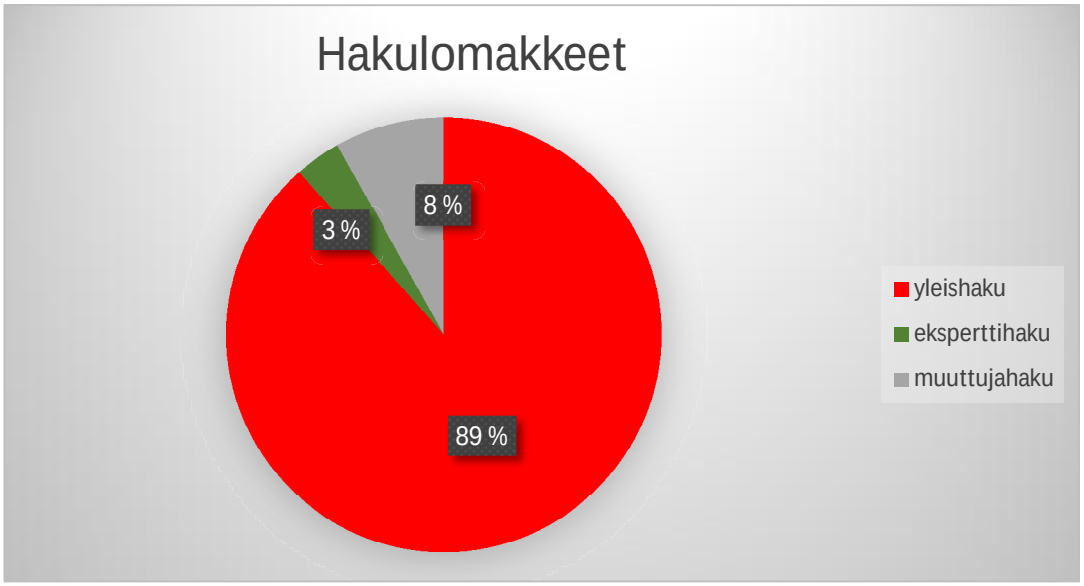
Lokin analysoinnin lisäksi tein *Ailassa* koehakuja päästäkseni tarkastelemaan lokista löytyvillä hakutermeillä saatavia tuloksia. Tein hakuja vapaasanahauista satunnaisotannalla, kymmenen prosenttia kaikista vapaasanahauista, poimimieni hakujen hakutermeillä. Näin selvitin näillä hauilla saadut tulosmäärät. Sen jälkeen otin vielä lähempään tarkasteluun nollatuloksia saaneet ja yli 45 tulosta tuottaneet vapaasanahaut, joiden hakutermeistä poimin esimerkkejä asiasanahakujen onnistumiseen vaikuttavista tekijöistä. Samalla tavalla tarkastelin niitä asiasanahakuja, joissa käytetyt hakutermit eivät lainkaan vastanneet asiasanastoja. Tämä hakutermien arviointi on tutkimukseni laadullinen osuus.

5 TULOKSET

Tässä luvussa esitellään lokianalyysin tulokset. Aliluvussa 5.1. vastataan tutkimuskysymykseen eri hakulomakkeiden käyttömääristä. Lisäksi esitellään yleishaun eri hakukenttien osuudet yleishakulomakkeen hauista vastaten tutkimuskysymykseen asiasana- ja vapaasanahaun määristä. Aliluvussa 5.2. vastataan kysymykseen asiasanojen osuudesta asiasanahaussa. Aliluku 5.3. käsittelee vapaasanahauista satunnaisotannalla valitulla näytteellä *Ailassa* tehdyn hakukokeilun tuloksia. Lisäksi esittelen ja arvioin siinä esimerkkien avulla vapaasana- ja asiasanahakujen onnistumiseen vaikuttavia tekijöitä. Aliluku 5.3. pohjautuu B -ryhmän tutkimuskysymyksiin (ks s.9) aiheita kuvailevan metatiedon ja järjestelmän hakuominaisuuksien vaikutuksista aineistojen löydettävyyteen.

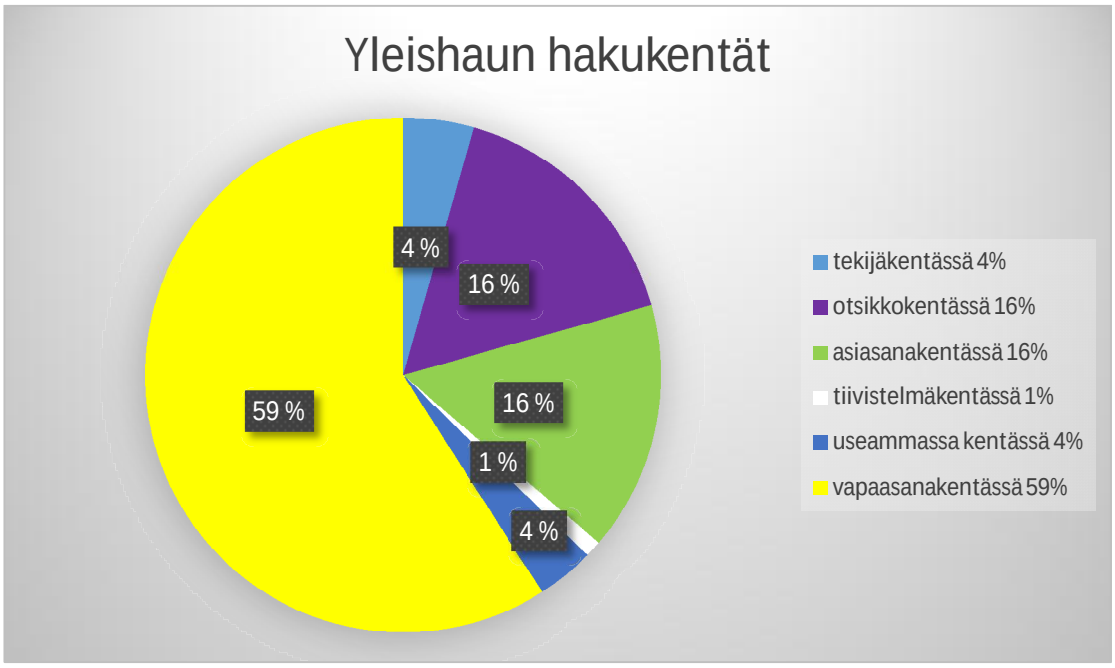
5.1Eri hakulomakkeiden ja yleishaun hakukenttien käyttö

Yleishakulomakkeella oli tehty 89 prosenttia kaikista hauista. Muuttujahakulomaketta oli käytetty 8 prosentissa ja eksperttihakulomaketta 3 prosentissa kaikista hauista. Ekspertti- ja muuttujahauista selvitin ainoastaan määrät. En siis tässä selvittänyt eri hakukenttien käytön osuutta muuttujahauissa. Hakulomakkeiden osuudet hakujen kokonaismäärästä näkyvät kuviossa 1.



KUVIO 1: Hakulomakkeiden osuudet

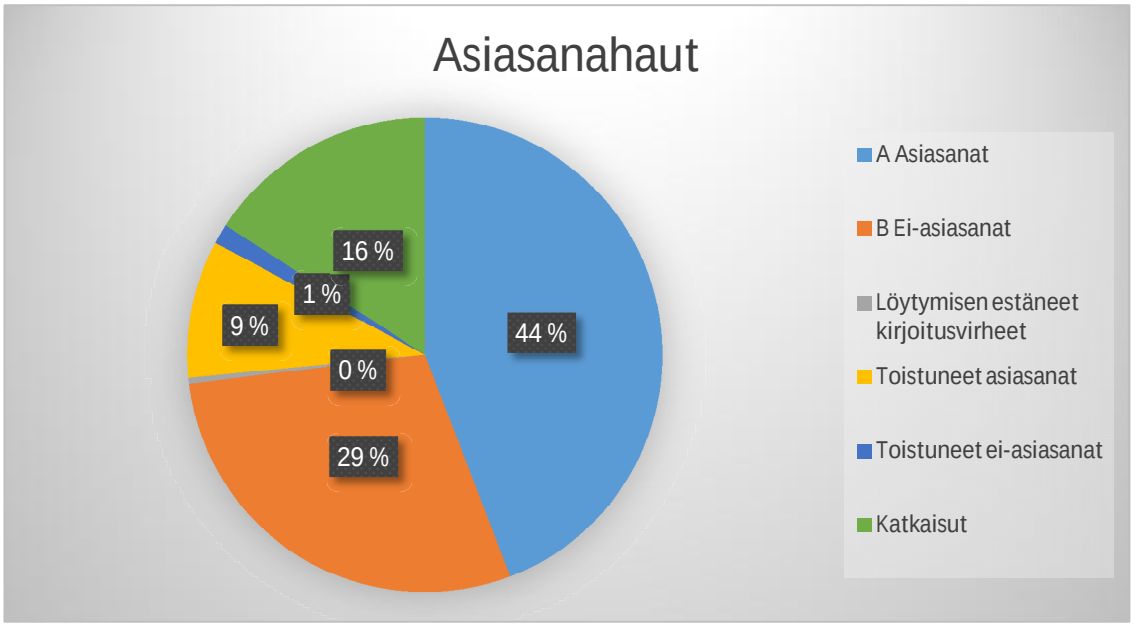
Selvitin eri hakukenttien käytön osuudet yleishakulomakkeella tehdyissä hauissa. Nämä näkyvät kuviossa 2. Yksinomaan vapaasanakenttää oli käytetty 59 prosentissa. Otsikko- ja asiasanakenttien käyttö oli yhtä yleistä: kummankin osuus oli 16 prosenttia. Tekijäkenttää käytettiin neljässä prosentissa hauista. Myös hakujen, joissa oli käytetty useampaa kenttää, osuus oli neljä prosenttia. Tiivistelmäkenttään kohdistettujen hakujen osuus oli yksi prosentti.



KUVIO 2: Yleishaun hakukentät

5.2Asiasanahaut

Asiasanakenttään kohdistettuja hakuja oli 396. Näistä 31 haussa oli asiasana-kenttään kirjoitettu kaksi tai useampia hakutermejä. Käytetyistä hakutermeistä kaikkiaan 226 oli asiasanastoista löytyviä asiasanoja. Ne siis olivat täsmälleen, taivutusmuodoissaan, YSA- tai ELLST- tesauruksista löytyviä sanoja. Hakutermejä, jotka eivät missään muodossa ole asiasanaston asiasanoja, oli käytetty 63 kertaa. Löytymisen estäneitä kirjoitusvirheitä hakutermeissä oli 2. Sanakatkaisun ansiostahan kirjoitusvirheet eivät estäisi löytymistä, jos ne sijaitsisivat katkaisun piiriin tulevassa sanan loppupäässä. Asiasanojen, ei-asiasanojen, katkaisujen ja toistojen osuudet näkyvät kuviossa 3.



KUVIO 3: Asiasanahaut. Asiasanojen, ei-asiasanojen ja kirjoitusvirheiden osuudet sekä toistojen ja hakua määrittäneiden sanakatkaisujen osuudet

Kaikista asiasanahauista asiasanastojen asiasanoja on siis käytetty 44 prosentissa ja sanoja, jotka eivät missään muodossa ole asiasanoja, 29 prosentissa. Toistot huomioin erikseen. Tämä tarkoittaa sellaisia hakuja, jotka on tehty samoin hakutermein useampaan, aineistossa yleensä kahteen tai kolmeen kertaan, mutta eri datatyyppi- tai kielisuodatuksin.

asiasanat 44 %	ei-asiasanat 29 %
toistuneet asiasanat 9 %	toistuneet ei-asiasanat 1 %
osuus asiasanahauista yhteensä 54%	osuus asiasanahauista yhteensä 30%

Lisäksi kuviossa 3 näkyy niiden hakujen osuus, joissa sanakatkaisu on vaikuttanut siihen, täsmääkö termi asiasanastoon. Se on 81 termiä eli 16 prosenttia.

5.2.1 Sanakatkaisu

Sanakatkaisun ansiosta *Ailan* hakujärjestelmä ”löysi” oikean muodon myös osalle niistä sanoista, jotka oli syötetty hakukenttään eri taivutusmuodossa kuin missä sana asiasanastossa on. Näin tapahtuu muun muassa monien suomen kielen monikkojen ja verbien infinitiivien kohdalla. Huomattava

on, että aineistostani ei saa täydellistä kuvaa hakujärjestelmän tekemästä stemmauksesta. Taivutusmuodoista näkyy siinä vain muutama ensimmäinen, ja sanakatkaisun toteutumista on arvioitava sanojen muodon arvioinnin ja koehakujen avulla. Tällä varauksella olen tarkastellut sanakatkaisuja muutamien esimerkkien valossa. Näistä ensimmäisenä *downshifting*

```
((((kdownshifting:(pos=1) OR ZKdownshifting:(wqf=2) OR ZKdownshift)
```

YSAn asiasana on *downshiftaus.* Järjestelmän tekemän sanakatkaisun ansiosta tulee tämän esimerkin hakuun mukaan myös asiansanaston asiasanaa vastaava sanamuoto. Huomio on kuitenkin vain teoreettinen, sillä *downshiftaus* -asiasanalla kuvailtua aineistoa ei *Ailasta* löydy. Selvittämättä jää tietysti myös se, onko hakijalla ollut mielessä suomen- vai englanninkielinen hakutermi.

Toisaalta on tapauksia, joissa sanan vartalo muuttuu muodosta toiseen siirryttäessä niin alussa, että ero säilyy katkaisusta huolimatta. Näin on käynyt esimerkiksi aineistossa olevassa *kulutuskriittisyys* -haussa. YSAn asiasana on *kulutuskritiikki*.

```
kkulutuskriittisyys:(pos=1) OR ZKkulutuskriittisyys OR ZKkulutuskriittisyy
```

Tässä siis ei sanakatkaisu näytä tavoittavan muotoa, joka yhdistäisi hakijan käyttämän termin asiasanaston vastaavaan. *Ailassa* ei kuitenaan ole myöskään *kulutuskritiikki* -asiasanalla kuvailtua aineistoa.

Toisena esimerkkinä mainittakoon *syrjäytynyt* ja *työtön,* joita on käytetty samassa haussa:

```
ksyrjäytynyt:(pos=1) OR ZKsyrjäytyny OR ZKsyrjäytynyt:(wqf=2)) AND (ktyötön:(pos=1) OR ZKtyötö OR ZKtyötön
```

YSAn asiasanat ovat monikkomuotoiset *syrjäytyneet* ja *työttömät.* Haku työttömät -asiasanalla tuottaa yhden tuloksen, kun taas *työtön* jää asiasanahaussa nollatulokselliseksi.

Tällaisia ”katkeamattomia” hakutermejä on kaiken kaikkiaan 12 kappaletta. Niiden osuus kokonaismäärästä on siis pieni. On huomattava, että kaikki järjestelmän generoimat sanamuodot eivät ole mukana aineistossa. Sanakatkaisun vaikutusta ei siis ole ollut tässä mahdollista kokonaan selvittää,

mutta kokonaishavaintoihin se ei vaikuta. Asia on kuitenkin huomionarvoinen kun kartoitetaan, millaisia asioita hakujärjestelmiä kehitettäessä on syytä huomioida.

Huomionarvoista on tässä myös se, että *Ailan* hakujärjestelmä ei pysty käsittelemään kaikkia käyttäjän itse tekemiä katkaisua. Esimerkiksi *lähisuhdeväkiv* -muotoon katkaistu *lähisuhdeväkivalta* asiasanakentässä tuottaa samat kaksi hakutulosta kuin sana katkaisemattomana, mutta *opiskel* -asiasanahaku ei tuota yhtään *opiskelu* -asiasanalla löytyvästä 36 tuloksesta, vaan nollatuloksen. Lisäksi katkaisu toimii eri tavoin vapaasanahaussa kuin asiasanahaussa. Vapaasanahaussa *opisk* tuottaa kahdeksan tulosta ja *opiskelu* 872.

Olen erikseen huomioinut myös sellaiset sanakatkaisut, jotka johdattavat useampaan asiasanaston asiasanaan. Näin on esimerkiksi hakijan itse katkaisemalla hakutermillä

```
kseksu:(pos=1) OR ZKseksu
```

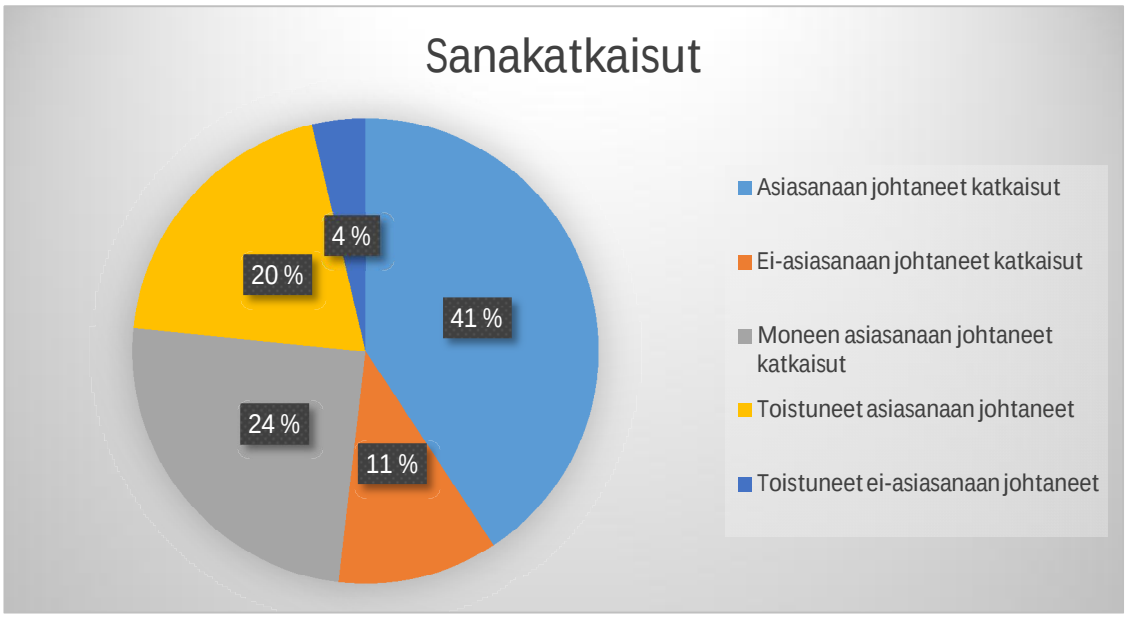
Tämä katkaisu voisi johtaa 22 eri YSAssa olevaan termiin.

Toisessa esimerkissä hakuterminä on *päihde:*

```
kpäihde:(pos=1) OR ZKpäihd
```

YSAssa on asiasana *päihteet,* jota tällä ei löytäisi, mutta siellä on myös 10 *päihde-* alkuista termiä, jotka ovat yhteen kirjoitettuja yhdyssanoja. Tällaisia moneen vaihtoehtoiseen asiasanaan johtavia katkaisuja on asiasanahakujen joukossa kaikkiaan 20. Edellä kuvatun *downshifting* -esimerkin kaltaisia asiasanaan johtavia katkaisuja on 33 kappaletta. *Syrjäytynyt* ja *työtön* -esimerkin kaltaisia katkaisuja, jotka eivät ole muuttaneet hakutermiä asiasanaksi, on kaikkiaan 9.

Myös katkaisujen kokonaismääriin ovat vaikuttaneet samanlaiset toistot kuin asiasanahakujen kokonaisuuteen eli samoin termein, mutta eri kieli-tai dokumenttityyppisuodattimin toistetut haut. Asiasanaan johtaneita katkaisuja ja niiden toistoja on yhteensä 49 kappaletta eli 61 prosenttia, ei-asiasanaan johtaneita toistoineen puolestaan 12 kappaletta eli 15 prosenttia sanakatkaisun määrittämistä hakutermeistä.

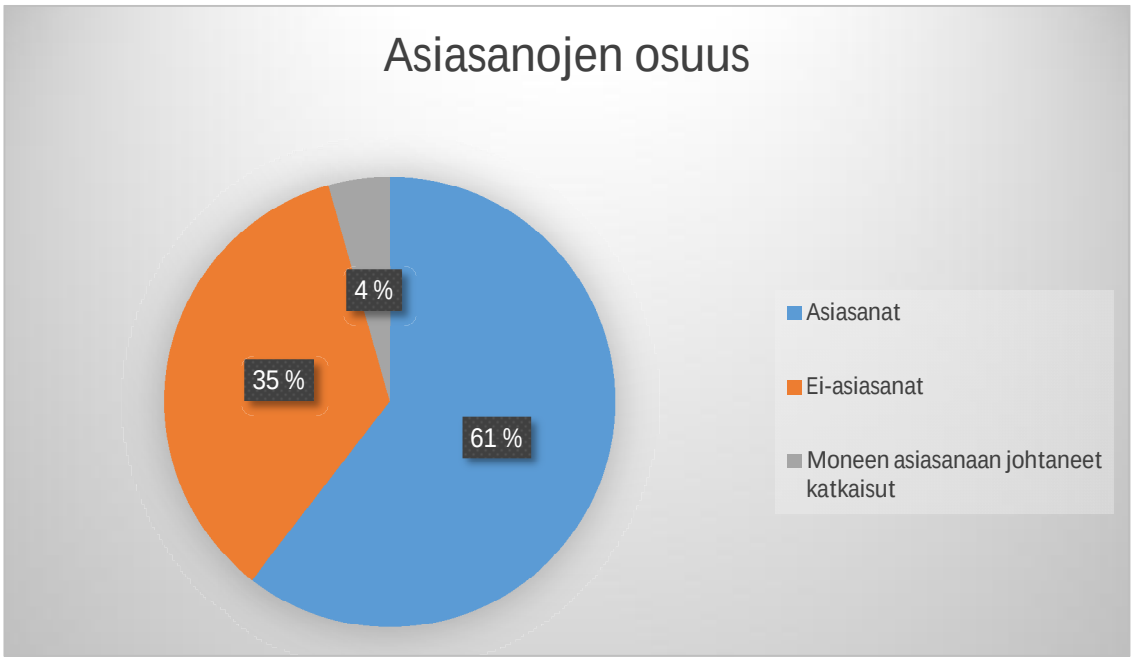


KUVIO 4: Asiasanastoon täsmäämiseen vaikuttaneet sanakatkaisut toistoineen.

5.2.2 Asiasanojen osuus asiasanahaussa käytetyistä termeistä

Näin voidaan asiasanojen osuus asiasanahauista määrittää lopuksi niin, että moneen asiasanaan johtaneet katkaisut jäävät vielä omaksi kategoriakseen.

Asiasanat	Ei-asiasanat	Moneen asiasanaan johtaneet katkaisut
275	160	20



KUVIO 5: Asiasanojen käytön osuus kaikista asiasanahauista.

Asiasanojen osuudeksi tulee näin jaotellen 61prosenttia kaikista asiasanakenttään kohdistetuissa hauissa käytetyistä hakutermeistä. Moneen asiasanaan johtaneiden sanakatkaisujen osuus, 4 prosenttia eli 20 hakutermiä, ei lopulta vaikuta suuresti kokonaisjakaumaan. Hakutermejä, jotka missään muodossa eivät ole asiasanoja, oli 160 kappaletta eli 35 prosenttia kaikista.

5.3. Muut aiheenmukaiset haut

5.3.1 Vapaasanahaut

Vapaasanahakujen joukosta satunnaisotannalla poimituista 149 hausta otin yli 45 hakutulosta tuottaneet 37 hakua kokeiluun, jossa vertasin termejä asiasanastoon ja tein vastaavat haut asiasanahakuina. Näin pääsin vertaamaan vapaasana- ja asiasanahakujen saantia. Yli 45 hakutulosta tuottaneissa hauissa seitsemässätoista käytetyt hakutermit olivat sellaisinaan myös YSAn tai ELLSTin asiasanoja. Nämä haut siis olisivat onnistuneet myös asiasanahakuina, jolloin tulosjoukkoon olisi oletettavasti saatu parempi tarkkuus.

TAULUKKO 1: vapaasanaotoksen yli 45 hakutulosta tuottaneet asiasanastoista löytyvät termit. Vasemmalla saanti vapaasanahaulla, oikealla saanti asiasanahaulla.

käytet/t/yt hakutermi/t	Vapaasana	Asiasana	käytet/t/yt hakuter-mi/t	Vapaasana	Asiasana
opetus	46	12	environmentalism	121	63
sosiaalinen media	46	11	power	153	66
Russia	51	7	workplace	186	19
hoito	52	8	eduskuntavaalit	187	79
työttömyys	59	45	perhe arvot	252	30
sosiaalinen tuki	60	36	perhe	306	104
perheet hyvinvointi	62	19	sosiaali	332	9
kokemukset	183	12	internet	882	234
kasvatus	68	44			

Seuraavaksi otin tarkasteluun nollatuloksen tuottaneet vapaasanahaut (liite 2), joita otoksessa oli 30. Näistä neljä nollatulosta oli johtunut hakutermin kirjoitusvirheestä tai vastaavasta ja jätin ne tarkastelun ulkopuolelle.

Jäljelle jäänyt 26 vapaasanahaun joukko on pieni nollatuloksellisuuden syiden laajempaan arviointiin. Se tarjoaa kuitenkin kiinnostavia havaintoja tilanteista, joissa haut eivät ole olleet toimivia – sekä viitteitä myös aiheista, joita todella ei ole kokoelmassa.

Näitä otoksen nollatuloksellisia hakuja on kiinnostavaa verrata myös Grossin ja Taylorin työhön, jossa muun tarkastelun ohessa listattiin Pittsburghin yliopistokirjaston asiakkaiden nollatuloksellisia hakuja (Gross & Taylor 2005, 219). Niiden joukossa, kaikkiaan 32 hakua, oli kuusi kirjoitusvirheel-

listä sekä kaksi sellaista, joissa asiakas oli epäonnistuneesti yrittänyt itse sanakatkaisua. Lopuista hauista useimmat vaikuttavat siltä, että ongelmaksi on muodostunut spesifisyyden taso. Vapaasanoja on käytetty samassa haussa useita ja niillä on yritetty muotoilla aihe varsin tarkasti saamatta osumaa lopulta mihinkään. Väljemmin muotoillen olisi tuloksia voinut löytyä. Tämä on toki pelkkä oma arvioni. Mielenkiintoisella tavalla FSD:n lokista poimimani nollatulokselliset vapaasanahaut kuitenkin muistuttavat Grossin ja Taylorin listaamia. Myös niistä monissa vaikuttaa ongelmana olevan aiheen astetta liian tarkka muotoilu. Esimerkkejä tästä ovat

suomalainen kulttuurinkuluttaja
henkilöauto
pappi kutsumus
ja *ulkoliikunta.*

Tarkkuuden tason lisäksi kompastuskivenä voi olla muoto. Siinä missä asiasanahaussa on tarpeen muotoilla haut indeksointikielen mukaisiksi, on vapaasanahaussa osuttava arkiston tarjoamia kuvailuja vastaaviin luonnollisen kielen ilmauksiin. Tästä esimerkkeinä mainittakoon

uusliberalismi
tehokkuusvaatimukset
parisuhdeväkivalta
internetin käyttökohteet
syrjäytynyt
väkivalta maskuliinisuus
flash eurobarometri
ja *salutogeeninen.*

Mukana on myös sellaisia aihetta kuvaavia termejä, joilla olettaisin aineistoa löytyvän. mikäli aiheesta aineistoa olisi. Esimerkiksi

hevonen
menetelmäoppaat
oil
*Tutkimisen taito (*joka on kirjan nimi)

ovat vapaasanahakuja, joiden nollatuloksellisuuden arvelen johtuvan siitä, ettei niillä haettua aineistoa arkiston kokoelmassa ole.

Nämä päätelmät ovat tietysti omia arvioitani. Nollatuloksellisten hakujen tarkastelu on kuitenkin mielenkiintoinen osa vapaasanahakujen analyysia, ja halusin sisällyttää myös sen työhöni pienenä lisänä. Jo tämä suppea otos antaa aihetta olettaa, että nollatuloksellisten vapaasanahakujen tarkastelu voi tarjota kiinnostavaa tietoa asiakkaiden tavoista käsitteellistää aiheita ja valita hakutermejä.

5.3.2 Asiasanakenttään ”eksyneet” vapaasanahaut

Tämän jälkeen tarkastelin niitä asiasanahakuja, joissa käytetyt hakutermit eivät vastanneet asiasanastoa eikä liioin järjestelmän tekemä sanakatkaisu muokannut niitä asiasanastoa vastaaviksi termeiksi. Näitä oli yhteensä 66 hakua ja 67 termiä. Useimmat olivat nollatuloksellisia, mutta eivät kaikki: ”*social policy*” tuottaa tietojärjestelmässä osuman sekä asiasanakentän merkkijonoihin ”*social*” että ”*policy*”. Tämä on sinänsä huomionarvoista järjestelmän asiasanahaun toimivuutta arvioitaessa.

TAULUKKO 2: asiasanahaussa esiintyneiden ei-asiasanojen piirteitä.

Sanamuoto	**Sanavalinta**	**Spesifisyys**	**Lyhenne**	**Puute sanastossa**
26	6	15	6	4

Määrittelin yllä kuvatut kategoriat eritelläkseni asiasanastoon täsmäämättömyyden syitä. Tämä on oma tulkintani aineistosta, asiasanastosta dokumentaatiokielenä ja luonnollisen kielen käytöstä. Termejä olisi siis mahdollista arvioida aivan toisinkin. Tulkintani kuitenkin tekee näkyväksi muutamia asiasanahaun ja asiasanastojen ongelmakohtia. Se perustuu pieneen otokseen enkä siksi esitä arvioita kunkin ongelman yleisyydestä.

Ongelma 1: sanavalinnat

Hakuterminä on käytetty luonnollista kieltä ”kääntämättä dokumentaatiokielelle”. Tähän kategoriaan olen ryhmitellyt ”melkein asiasanat”. Käytetyt termit ovat usein luonnollisessa kielessä dokumentaatiokielen termin synonyymeja. Mukana on asiasanaston sanaa arkikielisempiä synonyymeja, kuten *bussi,* ja arkikielessä tai tutkimuskielessä asiasanaston termin kanssa varsin tasavertaisesti

käytettyjä ilmauksia, kuten *omaishoiva, burnout* ja *työssäjaksaminen*. YSAn termit ovat *omaishoito, uupumus* ja *jaksaminen*, joista kahteen viime mainittuun johdatetaan muun muassa ohjaustermein *burnout, loppuunpalaminen* ja *työssä jaksaminen*.

Esimerkiksi *työolosuhteet, kuluttaminen* ja *kommunikointi* saattaisivat hyvin ollakin asiasanastossa, mutta eivät ole – YSAn termit ovat *työolot, kulutus* ja *tiedonvälitys*. *Kommunikaatiosta* neuvotaan YSAssa ohjaustermillä käyttämään termiä *tiedonvälitys* (YSA 2015). Luonnollisessa kielessä ja tutkimuskäsitteistössä ne merkitsevät hyvinkin eri asioita. *Social policy* taas antaa viitteitä monikieliseen asiasanastoon liittyvistä käsitteiden kääntämisen vaikeuksista. Suomessa käsite on yleinen ja sitä pidetään sosiaalipolitiikan oppialan englanninkielisenä nimenä. ELLST -asiasanastossa sitä ei kuitenkaan ole, vaan siinä käytetään ilmausta *social and welfare policy* (ELLST 2015).

Ongelma 2: sanamuodot

Myös tässä on käytetty luonnollista kieltä ”kääntämättä dokumentaatiokielelle” ja muotoja, joissa suomen kielen käyttäjillä on variaatiota ja asiasanastosta poikkeava ”väärä” kirjoitusasu on arkikielessä yleinen, kuten *aikakaus**i**lehti* ja *kiinteist**ö**välittäjä*. Joissain tapauksissa ei ole tiedetty, että asiasanastoissa käytetään yleensä monikkomuotoa. Niin on päädytty käyttämään esimerkiksi hakutermejä *koululainen, lukiolainen* ja *syrjäytynyt*.

> *Ongelma 2.1: lyhenteet*
> Hakulauseketta muotoiltaessa ei ole tiedetty, että asiasanastoissa ei käytetä lyhenteitä. Esimerkkeinä *DDR, PTSD* ja *EVA*.

Ongelma 3: ”liian spesifit” termit

Tämä on tieteellisen tiedonhaun kannalta kiintoisa ongelma. Monet otokseen tulleista ei-asiasanoista vaikuttavat olevan ikään kuin astetta spesifimpiä kuin asiasanasto. Tällaisista termeistä esimerkkeinä ovat *verkostoteoria, ilmastokasvatus, measures of democracy, yhteistutkiminen, erovanhemmuus* ja *hyvinvointitutkimus.*

Ongelma 4: Puutteet dokumentaatiokielessä

Tulkintani mukaan näiden termien pitäisi olla mukana asiasanastossa, mutta eivät ole. Näin esittäessäni ymmärrän tämän olevan oma näkemykseni asiasta. Myös edelliseen kohtaan kategorisoimani ”liian spesifit” termit voivat olla tulkittavissa osoituksiksi asiasanaston puutteellisuudesta. Toisaalta kullakin asiasanastolla on oma spesifisyyden tasonsa, jossa pitäytyminen ei kerro puutteellisuudesta, mutta kylläkin asiasanaston soveltuvuudesta eri aihepiirien kuvailuun. Tässä yhteydessä todettakoon myös, että yhteiskuntatieteiden alalla ei ole omaa suomenkielistä asiasanastoa.

Alakohtaisten asiasanastojen päivittäminen voi perustua esimerkiksi analyysiin alan julkaisuista. Tällöin asiasanastoon lisätään uusia termejä sillä perusteella, että ne esiintyvät riittävän usein asiasanaston päivitystyön pohjana käytetyissä alan teksteissä (ks. esim. Broughton 2006, 57–65). YSA pyrkii kattamaan varsin monia aloja ja sen päivitys perustuu paljolti sitä työssään käyttäviltä sisällönkuvailun tekijöiltä tuleviin ehdotuksiin (Kansalliskirjasto 2015). Tulkintani mukaan *äärioikeisto, klinikat* sekä aikakausien nimitykset ovat niin laajasti käytettyjä käsitteitä, että niiden sisällyttäminen YSAn tyyppiseen yleiseen asiasanastoon olisi perusteltua. Nyt ne eivät ole mukana edes ohjaustermeinä. Viimemainittu aikakausien nimeäminen on laaja kysymys. Otan sen tässä esiin asiasanahakuotoksessa mukana olleen *1960-luvun* johdosta. Se johti minut tarkastelemaan YSAn niin kutsuttuja vapaan indeksoinnin sanaryhmiä, joiden avulla voidaan dokumentaatiokieleen sisällyttää laajasti sellaisia monien eri aihepiirien termejä, joita ei ole otettu mukaan itse asiasanastoon. Aikakausia kuvaavia nimityksiä ei ole näissä mukana.

6 POHDINTA

Tässä luvussa arvioidaan tutkimuksen tuloksia yleisesti ja suhteessa alaluvussa 3.4. esitettyihin tutkimushypoteeseihin.

6.1 Tulokset suhteessa hypoteeseihin ja aiempaan tutkimukseen

Hypoteesi eri hakuvaihtoehtojen käytöstä, ”yleisintä on yksinkertainen vapaasanahaku”, pitää aineiston valossa paikkansa.

”Asiakkaat tapaavat käyttää yksinkertaista vasaraa tarjotun monipuolisen työkalupakin sijaan” eli valitsevat käytettävissä olevista hakuominaisuuksista yksinkertaisimmat (Antell & Jia 2008, McKay & Buchanan 2013) ei tämän tutkimuksen tuloksissa täysin pidä paikkaansa. *Ailan* käyttäjän ovat valinneet eri hakulomakkeita, käyttäneet yleishaun eri kenttiä ja yhdistelleet niitä. Yleishaun osuus hakulomakkeista on kuitenkin hyvin suuri, 89 prosenttia. Vapaasanahaun osuus yleishaun kentistä on suuri, mutta muitakin kenttiä, erityisesti asiasana-ja otsikkohakua, on käytetty melko usein: niillä kummallakin on 16 prosentin osuudet yleishakulomakkeen hauista.

Tässä yhteydessä tutkimukseni havaintoja voidaan rinnastaa myös Villén-Ruedan ja muiden tutkimukseen, jossa todettiin asiantuntevimpien käyttäjäryhmien suosivan otsikkohakua (Villén-Rueda et al 2007). Aineiston nimen tienneet – sen mahdollisesti ensin selailuhakuja tehdessään löytäneet – ovat tehneet otsikkohakuja *”FSD1234”*, *”työolobarometri”*, *”kuntasuomi 2004”*, ja *”erovanhemmuuskirjoitukset”* -tyyppisillä hakutermeillä. Otsikkokenttä voi tarkkuuden puolesta usein olla varteenotettava hakuvaihtoehto *Ailassa*, jossa otsikoiden käyttö vapaasanahaussa usein kasvattaa saannin satojen tulosten suuruudeksi ja romahduttaa tarkkuuden.

Hypoteesi aihetta kuvaavan metatiedon käytettävyydestä, ”asiasanahaku tuottaa vaikeuksia”, pitää paikkansa. Suoraan FSD:n aiheita kuvailevan metatiedon välineinä käyttämiä asiasanastoja YSAa ja ELLSTiä vastaavia termejä on käytetty 60 prosentissa kaikista asiasanahauista. Asiasanakentän hauista huomattava osuus on siis ollut enemmän tai vähemmän tuloksetonta – asiasanastoa vastaamattomat termit eivät useimmiten tuota tuloksia asiasanahaussa. Tässä tarkastelussa on huomioitu myös järjestelmän tekemien sanakatkaisujen vaikutus.

Muotoilin aluksi hypoteesin myös aihetta kuvaavan metatiedon laadun merkityksestä: ”lokiaineisto kertoo myös dokumentaatiokielten laadusta ja sopivuudesta.” Tarkasteluni antaa tästä joitain viitteitä, mutta kattavampi arviointi olisi uuden tutkimuksen aihe. Asiasanakentässä käytetyistä hakutermeistä useat johtavat pohtimaan nimensä mukaisesti yleisen YSAn tarkkuuden tasoa suhteessa kuvailtuihin aineistoihin. Eräät englanninkieliset termit taas herättävät pohtimaan monikielisten tesaurusten käytettävyyskysymyksiä. Näitä asioita ei ole kuitenkaan tässä työssä systemaattisesti tarkasteltu.

6.2. Havaintojen merkitys ja tutkimuksen arviointi

Onko tällä tutkimuksella jotain uutta kerrottavaa asiakkaiden tekemistä hauista? Aiemman tutkimuksen pohjalta esitetyt hypoteesit näyttivät pitävän paikkansa ja havainnot olevan varsin samankaltaisia kuin taustalukemistona käytetyissä tutkimuksissa.

Jo se, että tässä tutkimuksessa saatiin varsin samansuuntaisia tuloksia kuin eri kontekstiin, kirjastoihin, sijoittuvissa tutkimuksissa, on huomionarvoista ja kiinnostavaa. Yhteiskuntatieteellisen tietoarkiston asiakaskunnan hakutavat ovat myös yleisesti mielenkiintoisia. Voidaan olettaa, että tutkija- ja opiskelijavaltainen asiakaskunta edustaa käyttäjäryhmää, joka hakee ja käyttää myös esimerkiksi tieteellisten kirjastojen aineistoja. Heidän hakutapansa voivat olla muidenkin organisaatioiden kannalta mielenkiintoisia ja tulokset jossain määrin yleistettävissä. En tässä kuitenkaan väitä, etteivätkö samat asiakkaat toimisi myös eri tavoin eri järjestelmissä riippuen kulloisenkin järjestelmän ominaisuuksista, opastuksesta, vaihtelevista tiedontarpeistaan, tilannetekijöistä ynnä muusta. Arkistomaailmaan sijoittuvaa lokitutkimukseen perustuvaa hakujen analyysia taas on toistaiseksi tehty vähän, tutkimusaineistoarkistojen piirissä ei ilmeisesti juuri lainkaan. Työni herättääkin uusia kysymyksiä ja tutkimusideoita. Niitä käsittelen seuraavassa, loppuluvussa.

Ennen sitä tahdon kuitenkin myös esittää muutamia arvioita itse tutkimuksen tekemisestä, kysymyksenasettelusta ja aineiston käsittelystä. Tämä työ tutustutti minut aineistoon ja tutkimusmenetelmään, jollaisia en ennen ollut tuntenut enkä käyttänyt. Painotin kysymyksenasettelussa minua aihetta kuvailevaan metatietoon liittyvän kiinnostukseni johdosta erityisesti kiinnostavaa teemaa: aiheenmukaisessa haussa käytettyjä termejä. Tämä teki tutkimuksestani hieman erilaisen kuin sellaiset lokianalyysit, joissa on pyritty pureutumaan yksittäisten asiakkaiden hakukäyttäytymiseen esimerkiksi hakusessioiden kartoittamisella. Sessioita tutkimalla olisin saanut tietoa esimerkiksi siitä, kuinka hakuja osataan, jos osataan, muokata. Edelleen, jos olisin voinut selvittää hakuja seu-

ranneet aineistotilaukset samaan tapaan kuin Huurninkin ja muiden AV-arkiston asiakkaiden hakuja selvittäneessä työssä (Huurnink et al 2010 a), olisin pystynyt jonkin verran arvioimaan hakujen onnistumista.

Olen tyytyväinen valitsemaani lähestymistapaan ja painotuksiin. Aloittaessani tutkimuksen olin tutustunut lokitutkimusmenetelmän hyvistä ja huonoista puolista käytyyn keskusteluun, jossa on usein mainittu käyttäjän intentioiden selvittämisen mahdottomuus lokitutkimuksen keinoin. Työtä tehdessäni huomasin, kuuinka mielenkiintoista voisi olla esimerkiksi käyttäjien haastatteleminen yhdistettynä lokista saatuihin havaintoihin. Käytettävyystutkimuksen menetelmät, joiden avulla esimerkiksi Riipinen (Riipinen 2014) on gradussaan selvittänyt Yhteiskuntatieteellisen tietoarkiston asiakkaiden näkemyksiä *Ailaa* edeltäneen hakujärjestelmän toimivuudesta, auttaisivat vastaamaan uusiin kysymyksiin, joihin lokiaineisto ei voi tarjota vastauksia.

Toisaalta minulle valkeni myös selvästi lokitutkimusmenetelmän vahvuus, joka tekee siitä kiehtovan ja mielenkiintoisen keinon tarkastella hakukäyttäytymistä. Lokista näkyy sellainen hakukäyttäytyminen, johon tietoisuus hakua havainnoivasta tutkijasta ei ole vaikuttanut. Pääsin selvästi toteamaan sen, miten toimiva väline lokitutkimus on hakutapojen tarkastelussa: *”one of the most reliable tools for observing search behaviour in the wild”* (McKay&Buchanan 2013).

Käyttäjien tietämättömyys tutkittavina olemisesta voi tuoda lokitutkimukseen myös tutkimuseettisiä kysymyksiä. Koska en selvittänyt yksittäisten käyttäjien sessioita, oli tässä tutkimuksessa tilanne niiden suhteen yksinkertaisempi kuin esimerkiksi IP -tunnistusta käyttävissä tutkimuksissa. Myös tähän tutkimukseen ryhtyessäni sitouduin pitämään aineiston salassa.

Työhön ryhtyessäni en myöskään tiennyt enkä osannut arvioida kuhunkin tarkasteluun tarvittavaa työmäärää ja aikaa. Vapaasana- ja asiasanahaussa käytettyjen termien kartoitus ja erityisesti hakujen toistaminen tulosten tarkastelemiseksi laajemmin olisi ollut kiinnostavaa myös tässä toteutunutta laajemmassa mittakaavassa. Näin olisin voinut saada kattavampia tuloksia, mutta aikaa olisi tarvittu enemmän kuin gradututkimuksen verran. Niin jouduin myös luopumaan aluksi suunnittelemastani vapaasanahaun tulosten tarkastelusta sen suhteen, kuinka usein vapaasanahaut tuottavat osumia asiasanoitukseen (vrt. Gross & Taylor 2005).

Lisäksi huomasin, että taulukkolaskentaohjelmassa tekemäni aineiston järjestäminen ja tarkastelu jätti osin sijaa inhimillisille virheille. Tämä huomio koski erityisesti sen laskemista, kuinka monessa

yleishakulomakkeen haussa oli käytetty useampaa kuin yhtä hakukenttää. Samoin sanakatkaisun vaikutus asiasanahakuun oli hieman olettamaani hankalammin selvitettävissä. Uskon kuitenkin saaneeni laskettua oikeat prosenttiosuudet ja esittämäni tulosten kokonaiskuvan olevan luotettava.

LOPUKSI

Aineistojen löydettävyys on olennainen asia avoimen datan politiikan varsinaisen toteutumisen kannalta. Yhteiskuntatieteellinen tietoarkisto on ottanut datan kannalta merkittävän askeleen asettaessaan aineistot *Aila*-portaalin kautta suoraan saataville: asiakkaat voivat nyt ladata löytämiään aineistoja suoraan omaan käyttöönsä.

Tämä tutkimus antaa kuvaa siitä, millä tavoin asiakkaat hakevat aineistoja *Ailasta* ja viitteitä siitä, kuinka he ovat niiden löytämisessä onnistuneet. Havaintojeni pohjalta olen ehdottanut Yhteiskuntatieteelliselle tietoarkistolle, jossa muutenkin on suunniteltu uusien hakuominaisuuksien lisäämistä *Ailaan* ja järjestelmän kehittämistä, seuraavia asioita:

1. Asiakkaiden tiedonhakutavat ovat tämän tutkimuksen valossa sellaisia, että suuria ja epätarkkoja saanteja on paljon. Siksikin ensiarvoisen tärkeää on luoda mahdollisuus palata rajaamaan hakutuloksia.

2. Asiakkaita tulisi ohjata nykyistä paremmin aiheita kuvailevan metatiedon käyttöön haussa. Erityisesti asiasanastojen käyttöön asiasanahaussa tarvitaan opastusta. Jos mahdollista, myös järjestelmää olisi muutettava sellaiseksi, että se tukee asiasanahakua. Tesauruksen rakennetta voitaisiin tehdä ymmärrettäväksi ja haussa hyödynnettäväksi tuomalla näkyviin ohjaustermejä, suppeampia termejä ja laajempia termejä. Muutosten jälkeen on mahdollista myös tehdä uudestaan tässä työssä tekemäni kaltainen tarkastelu asiasanahausta.

3. En ole tutkinut sitä, millä tavalla asiakkaat ovat hauissaan hyödyntäneet luokitusta. Yhteiskuntatieteellisen tietoarkiston aineistoja ja sen käytössä olevia, huolellisesti laaditun, kattavan ja käyttökelpoisen oloisia luokitustyökaluja katsellessani olen kuitenkin arvellut, että luokitukseen perustuva selailu voisi olla monissa tapauksissa varsin toimiva hakutapa Yh-

teiskuntatieteellisen tietoarkiston aineistoilla. Luokitukseen perustuvan haun mahdollisuutta voisi lisätä myös *Ailaan*.

4. *Ailan* eksperttihaku ei vaikuta saaneen ylipäätään paljoa käyttäjiä. Käyttäjät eivät liioin useinkaan ole hyödyntäneet sitä *”advanced search”* -tyyppisesti. Eksperttihaun käyttökelpoisuutta voisi miettiä. Samoin voisi pohtia, olisiko asiakkaita mahdollista saada käyttämään *”advanced search”* -tarpeisiin jo ennestään olemassa olevaa Nesstar-hakua. Nesstarissa on paljon hakumahdollisuuksia, mutta se on tähän mennessä ollut vähän käytetty ja käyttöliittymä nykymuodossaan käyttäjäepäystävällinen.

Siinä, miten asiakkaat osaavat hyödyntää aiheita kuvailevaa metatietoa haussa, riittäisi monipuolista tutkittavaa informaatiotutkimuksen alalle. Tässä työssä tehdyt havainnot asiasanahaussa käytetyistä hakutermeistä sekä vapaasana- ja asiasanahakujen saantien vertailut antavat viitteitä siitä, että niin tiedonhakukäyttäytymisestä, asiasanojen toimivuudesta hakuelementteinä kuin monista muistakin aiheista on lokitutkimuksen keinoin saatavissa paljon tietoa.

Aineistoarkistot ovat informaatiotutkimukselle huomionarvoinen ja toistaiseksi vähän tutkittu tutkimusalue. Niiden käyttö ja yhteiskunnallinen merkitys varmasti lisääntyy tulevaisuudessa avoin data-politiikan myötä, ja informaatioammattilaisten osaamista tarvitaan varmasti aineistoarkistoissa entistä enemmän.

LÄHTEET

Antell, Karen & Huang, Jia (2008): Subject Searching Success Transaction Logs, Patron Perceptions, and Implications for Library Instruction. In *Reference & User Services Quarterly*, 48(1).

Asunka, Stephen, Chae, Hui Soo, Hughes, Brian & Natriello, Gary (2009): Understanding Academic Information Seeking Habits through Analysis of Web Server Log Files; The Case of the Teachers College Server Website. In *Journal of Academic Librarianship*, 35 (1).

Borg, Sami: Tietoarkisto tukee tutkimusaineistojen jatkokäyttöä. Julkaisussa *Tieteessä tapahtuu* 3/2011, s. 1−2.

Borg, Sami & Kuula, Arja (2007): *Julkisrahoitteisen tutkimusdatan avoin saatavuus ja elinkaari. Valmisteluraportti OECD:n datasuosituksen toimeenpanomahdollisuuksista Suomessa.* Yhteiskuntatieteellinen tietoarkisto, Tampere.

Yhteiskuntatieteellinen tietoarkisto, Tampere. Borst, Timo (2012): Usage and Impact of Controlled Vocabularies in a Subject Repository for Indexing and Retrieval. In *Liber Quarterly* 21 (3/4).

Broughton, Vanda (2006): *Essential thesaurus construction.* Facet Publishing, London.

Chowdury, G G & Chowdury, Sudatta (2007): *Organizing information from the shelf to the web.* Facet Publishing, London.

DDI (2015): Data Documentation Initiative. http://www.ddialliance.org. Viitattu 2.3. 2015.

ELLST (2015): European Language Social Science Thesaurus. http://elsst.esds.ac.uk/Home.aspx. Viitattu 30.1.2015.

Euroopan Unioni (2012): Komission suositus tieteellisen tiedon saatavuudesta ja säilyttämisestä. Julkaisussa *Euroopan Unionin virallinen lehti* ((2012/417/EU).

Gross, Tina & Taylor, Arlene G. (2005): What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results. In *College & Research Libraries*, 66 (3).

Haynes, David (2004): *Metadata for information management and retrieval.* Facet Publishing, London.

Heinonen, Matti, atk-erikoistutkija (Yhteiskuntatieteellinen tietoarkisto): Sähköpostitiedonanto 4.11. 2015.

Hider, Philip ((2012): *Information resource description. Creating and managing metadata.* Facet Publishing, London.

Hildreth, Charles R (1997): The Use and Understanding of Keyword Searching in a [sic] University Online Catalog. In *Information Technology and Libraries* 16 (6).

Huurnink, Bouke, Snoek, Gees G.M., de Rijke, Maarten & Smeulders, Arnold W.M (2010a): Today's and Tomorrow's Retrieval Practice in the Audiovisual Archive. Konferenssipaperi, CIVR 10, Xi'an, Kiina, 2010. In *ACM Conference Proceedings*.

Huurnink, Bouke, Hollink, Laura, van den Heuvel, Wietske & de Rijke, Maarten (2010b): Search Behavior of Media Professionals at an Audiovisual Archive: A Transaction Log Analysis. In *Journal of the American Society for Information Science and Technology*, 61 (6).

Huuskonen, Saila & Vakkari, Pertti (2007): Students' search process and outcome in Medline in writing an essay for a class on evidence-based medicine. In *Journal of Documentation* 64 (2).

IFLA (2015): IFLA Code of Ethics for Librarians and other Information Workers. http://www.ifla.org/news/ifla-code-of-ethics-for-librarians-and-other-information-workers-full-version. IFLA International Federation of Library Associations. Viitattu 9.3.2015.

ISO 25964-1 (2011): The International Organization for Standardization: Thesauri for information retrieval. https://www.iso.org/obp/ui/#iso:std:iso:25964:-1:ed-1:v1:en. Viitattu 1.2.2015.

Jansen, Bernard J. (2006): Search log analysis: What it is, what's been done, how to do it. In *Library & Information Science Research* 28.

Jansen, Bernard J. (2009): The Methodology of Search Log Analysis. In Jansen, Bernard J, Spink, Amanda &Taksai, Isak (2009): *Handbook of research on web log analysis*. IGI Global.

Järvelin, Kalervo & Sormunen, Eero (2006): Dokumentit kateissa? Tiedon tallennus ja haku avuksi. Teoksessa *Tiedon tie. Johdatus informaatiotutkimukseen.* BTJ Kirjastopalvelu Oy.

Kansalliskirjasto (2015): YSA. Yleinen suomalainen asiasanasto. http://www.kansalliskirjasto.fi/kirjastoala/asiasanastot/ysa.html. Viitattu 30.1. 2015

Kumpulainen, Sanna (2013): *Task-based information access in molecular medicine: task performance, barriers, and searching within a heterogeneous information environment.* Tampere University Press, Tampere.

Larson, Ray R. (1991): The Decline of Subject Searching: Long-Term Trends and Patterns of Index Use in an Online Catalog. In *Journal of the American Society for Information Science and Technology*, 42 (3).

Lau, Eng Pwey (2006): In search of query patterns: a case study of a university OPAC. In *Information Processing & Management*, 42 (5).

Mukala, Seija (2005): *Hyvin toimivien hakulausekkeiden muotoilu ja hakujen onnistumiseen vaikuttavat tekijät täys- ja osittaistäsmäyttävässä hakujärjestelmässä.* Tampereen yliopisto, Tampere. Pro gradu -työ.

McKay, Dana & Buchanan, George (2013): Boxing clever: how searchers use and adapt to a one-box library search. Konferenssipaperi, OzCHI 13, Adelaide, Australia 2013. In *ACM Conference Proceedings*.

Niu, Xi & Hemmiger, Bradley M (2010): Beyond Text Querying and Ranking List: How People Are Searching Through Faceted Catalogs in Two Library Environments. In *Proceedings of the American Society for Information Science and Technology 47 (1)*.

Opetus- ja kulttuuriministeriö (2015): Avoin tiede ja tutkimus 2014–2017 -hankkeen verkkosivut. http://avointiede.fi, Viitattu 9.3.2015.

Opetus-ja kulttuuriministeriö (2015): Tutkimuksen tietoaineistot -hankkeen verkkosivut. https:www.tdata.fi/tta-ja-yhteistyo. Viitattu 9.3.2015.

Pynnönen, Marjaana (2000): *Lokitiedostot tiedonhakututkimuksessa ja käyttäjien toiminta tiedonhaussa: esimerkkinä Aamulehden internet-arkisto*. Tampereen yliopisto, Tampere. Pro gradu -työ.

Riipinen, Juha (2014): *Yhteiskuntatieteellisen tietoarkiston verkkosivujen käytettävyys noviisikäyttäjien näkökulmasta.* Tampereen yliopisto, Tampere. Pro gradu -työ.

Sihvonen, Anni & Vakkari, Pertti: Subject knowledge improves interactive query expansion assisted by a Thesaurus. In *Journal of Documentation* 60 (6), 2004.

Suominen, Vesa, Saarti, Jarmo ja Tuomi, Pirjo (2009): *Bibliografinen valvonta. Johdatus luettteloinnin ja sisällönkuvailun menetelmiin.* BTJ Kustannus, Helsinki.

Syväjärvi, Antti (2014): Tutkimuksen ja tutkimuslaitosten datapolitiikka uudistusten äärellä. Julkaisussa *Hallinnon tutkimus* 33 (1).

Valtiovarainministeriö (2014): *Avoimen tiedon ohjelma*. http://www.vm.fi/vm/fi/05_hankkeet/02381_avoin_tieto/index.jsp ,Viitattu 21.10.2014.

Vakkari, Pertti (2006): Tiedonhankinnan tukeminen ja informaatiotutkimus. Teoksessa *Tiedon tie. Johdatus informaatiotutkimukseen.* BTJ Kirjastopalvelu Oy.

Villén-Rueda, Luis, Senso, Jose A. and de Moya-Anegón, Félix (2007): The Use of OPAC in a Large Academic Library: A Transactional Log Analysis Study of Subject Searching. In *Journal of Academic Librarianship* 33 (3).

Wofram, Diemar, Wang, Peiling & Zhang, Jin (2009): Identifying Web Search Session Patterns Using Cluster Analysis: A Comparison of three Search Environments. In *Journal of the American Society for Information Science and Technology* 60 (5).

Xapian (2015): Xapian Search Engine Library. www.xapian.org. Viitattu 11.3.2015

Yhteiskuntatieteellinen tietoarkisto: *Yhteiskuntatieteellisen tietoarkiston arkistonmuodostussuunnitelma.* http://www.fsd.uta.fi/fi/hallinto/asiakirjat/AMS/ams_index.html. Viitattu 9.1.2015.

Yhteiskuntatieteellinen tietoarkisto (2014): *Toimintakertomus 2013*. Yhteiskuntatieteellinen tietoarkisto.

Yhteiskuntatieteellinen tietoarkisto (2015): *Aila*-portaalin hakuohje. https://services.fsd.uta.fi/catalogue/search?tab=instructions. Viitattu 30.1. 2015.

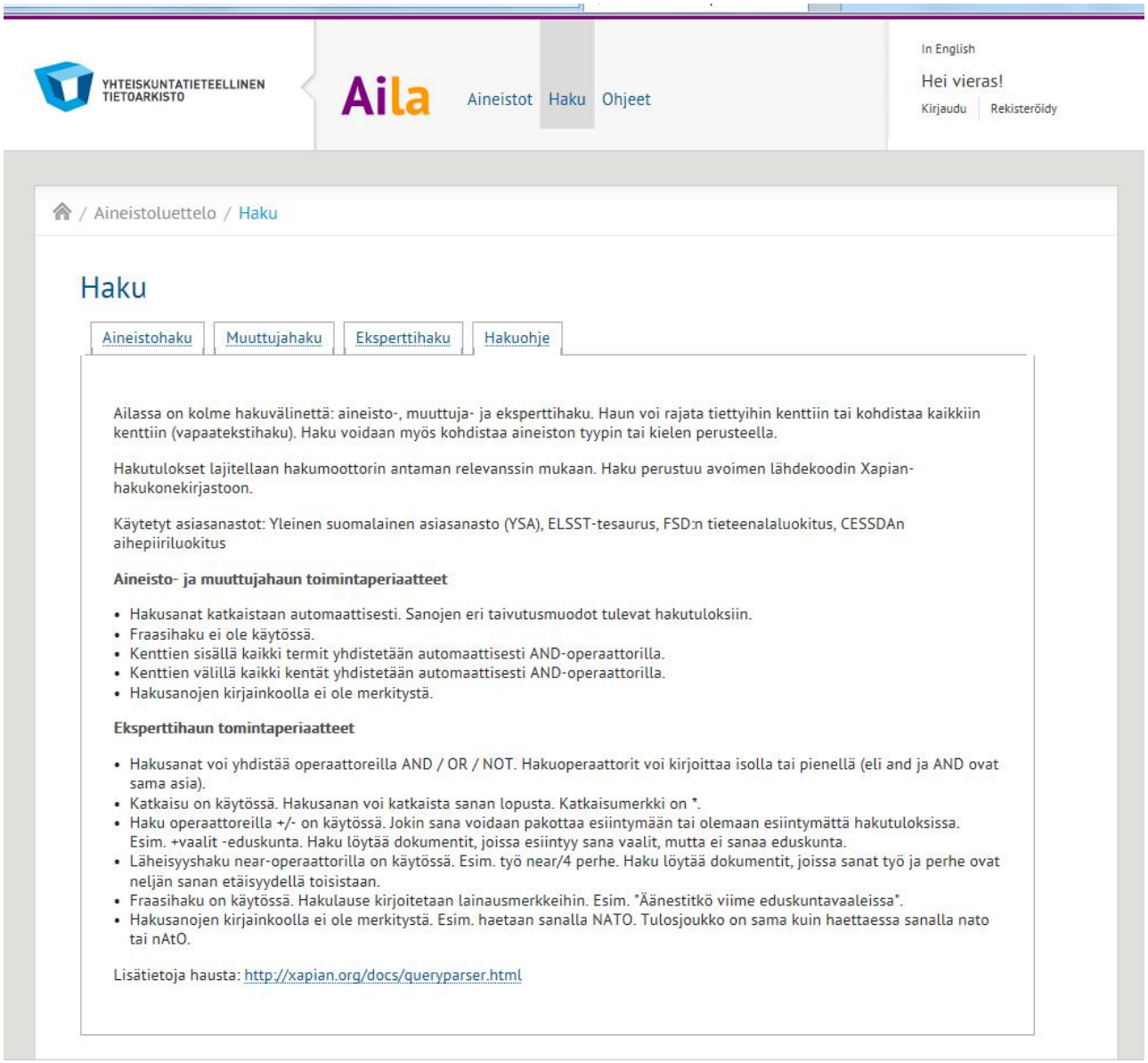
Yhteiskuntatieteellinen tietoarkisto (2015): *Aila lisäsi aineistojen käyttöä.* http://www.fsd.uta.fi/fi/ajankohtaista/tiedotteet/tiedote323.html. Viitattu 1.2.2015.

YSA (2015): Yleinen suomalainen asiasanasto. http://vesa.lib.helsinki.fi/ysa/index.html. Viitattu 30.1.2015.

Yu, Holly & Young, Margo (2004): The Impact of Web Search Engines on Subject Searching in OPAC. In *Information Technology and Libraries* 12/2004.

Zavalina, Oksana L. (2011): Contextual Metadata in Digital Aggregations: Applications of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval. In *Journal of Library Metadata* 11 (3-4).

LIITTEET



LIITE 1: *Ailan* hakuohje. 30.1. 2015.